DOCUMENT RESUME

ED 205 193 IR 009 497

AUTHOR Zachos, Paul
TITLE A Handbook for Systematic Evaluation in ESEA Title IV-B.
INSTITUTION New York State Education Dept., Albany. Bureau of School Libraries.
PUB DATE Apr 80
NOTE 86p.

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS *Educational Resources; *Evaluation Methods; Federal Aid; *Federal Programs; Guidelines; Models; Private Schools; *Program Evaluation; School Districts; *School Libraries
IDENTIFIERS *Elementary Secondary Education Act

ABSTRACT

Based on a survey of individuals responsible for Elementary and Secondary Education Act (ESEA) Title IV-B evaluations at the local level and an examination of IV-B applications, this handbook has been prepared to provide assistance to New York school districts and nonprofit private schools in the coordination and evaluation of school library and learning resources sub-projects. The content and organization allow the inexperienced individual to sequentially approach the text and the reader with more familiarity to select only those topics relevant to his needs. The handbook contains a discussion of some of the problems related to IV-B evaluations, an introduction to basic principles of education evaluation, a typology to assist in determination of evaluation appropriate to a given sub-project, an outline of steps to follow in planning the evaluation, and a review and sampling of evaluation procedures particularly relevant to IV-B sub-projects. Suggestions for selecting from commercially produced instruments are appended and a bibliography lists nine publications. (RAA)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED205193

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

A Handbook for Systematic Evaluation in ESEA Title IV-B

The University of the State of New York
THE STATE EDUCATION DEPARTMENT
Bureau of School Libraries
Bureau for Program Planning and Evaluation
Albany, New York 12234

April 1980

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Robert Trombley

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

IR009497

THE UNIVERSITY OF THE STATE OF NEW YORK

Regents of The University (with years when terms expire)

1981	Theodore M. Black, A.B., Litt.D., LL.D., Pd.D., D.C.L., L.H.D. Chancellor	Sands Point
1988	Willard A. Genrich, LL.B., L.H.D., LL.D., Litt.D. Vice Chancellor	Buffalo
1986	Kenneth B. Clark, A.B., M.S., Ph.D., LL.D., L.H.D., D.Sc.	Hastings on Hudson
1983	Harold E. Newcomb, B.A.	Owego
1982	Emlyn I. Griffith, A.B., J.D.	Rome
1983	Mary Alice Kendall, B.S.	Rochester
1984	Jorge L. Batista, B.A., J.D., LL.D.	Bronx
1982	Louis E. Yavner, LL.B.	New York
1986	Laura Bradley Chodos, B.A., M.A.	Vischer Ferry
1980	Martin C. Barell, B.A., I.A., LL.B.	Kings Point
1981	Joseph R. Bongiorno, B.S., M.D.	Brooklyn
1984	Louise P. Matteoni, B.A., M.A., Ph.D.	Bayside
1981	J. Edward Meyer, B.A., LL.B.	Chappaqua
1985	Arlene B. Reed-Delaney, B.A., M.D.	Albany
1987	R. Carlos Carballada, B.S.	Arcade

President of The University and Commissioner of Education
Gordon M. Ambach

Executive Deputy Commissioner of Education
Joseph J. Blaney

Deputy Commissioner for Elementary, Secondary and Continuing Education
Robert R. Spillane

Assistant Commissioner for General Education and Curricular Services
Maria Ramirez

Director, Division of Humanities and Arts
Charles J. Trupia

Chief, Bureau of School Libraries
Lore Scurrah

Assistant to the Commissioner for Policy Analysis
Claudio R. Prieto

Chief, Bureau for Program Planning and Evaluation
Howard Berkun

FOREWORD

This handbook has been prepared to provide assistance to those individuals in local school districts and nonprofit private schools who are responsible for the coordination and evaluation of projects funded under ESEA Title IV-B.

The handbook contains:

- a discussion of some of the problems related to IV-B evaluations
- an introduction to basic principles of educational evaluation
- a typology which will assist readers in determining what type of evaluation is appropriate for their sub-projects
- an outline of steps to follow in planning a IV-B evaluation, and
- a review and sampling of evaluation procedures, which are particularly relevant to Title IV-B sub-projects.

The scope of the handbook was influenced by the findings of a survey conducted by the Bureau for Program Planning and Evaluation in the spring of 1979 in which a sample of people responsible for Title IV-B sub-projects across New York State were asked to describe their evaluation methods, comment on the strengths and weaknesses of these methods and provide statements of their concerns about IV-B evaluation and recommendations related to the evaluation of Title IV-B sub-projects.

The content and organization of this handbook have been influenced by the findings of this survey as well as by an examination of a sample of IV-B applications. These findings support the belief that persons responsible for IV-B sub-projects have had varying degrees of formal training in educational evaluation. Readers who have more familiarity with the area of evaluation may choose to use the handbook by referring to those topics which they feel are relevant to their information needs, while others may choose to approach the material sequentially from start to finish. The contents have been written and arranged to be approached in either manner.

An attempt was made to present techniques and suggestions which minimize the time and effort needed to conduct a valid and useful evaluation. The present text should be useful in planning for Title IV-B sub-projects as well as suggesting ways to design and conduct evaluations of these sub-projects.

This document was prepared by Paul Zachos with the assistance of the staffs of the Bureau for Program Planning and Evaluation, and the Bureau for School Libraries.

HOW TO USE THIS HANDBOOK

This Handbook may be used in two ways: the text may be approached sequentially or the Table of Contents may be used as a guide to topics of interest. The following overview will help users of this handbook to identify portions which will be of greatest value to them:

Chapter 1. THE CONTEXT OF TITLE IV-B EVALUATION

This chapter reviews general characteristics of Title IV-B and of the New York State Title IV-B program. It is intended primarily for those who are not already familiar with Title IV-B sub-projects.

Chapter 2. OBJECTIVES AND EVALUATION IN IV-B SUB-PROJECTS

This chapter provides a rationale for the type of evaluation which is required by New York State's Title IV-B program; i.e., objectives-based evaluation. An INPUT-PROCESS-OUTCOME model of educational programs is presented in this chapter as well as a distinction between two types of objectives in educational programs -instrumental objectives and performance objectives. These concepts are fundamental to understanding the remainder of the evaluation model presented in this Handbook.

Chapter 3. A TYPOLOGY OF IV-B SUB-PROJECTS

This chapter classifies IV-B sub-projects into three types. Different types of evaluation are appropriate to each of these three types. This chapter will allow readers to classify their sub-projects, and will refer them to those portions of chapters 6 through 8 which are appropriate to their needs.

Chapter 4. A SEVEN-STEP MODEL FOR PLANNING EVALUATION OF IV-B SUB-PROJECTS

These steps, first presented in outline form, provide a guide through the evaluation design and planning process, including the development and refinement of the evaluation instrument. More detailed information regarding each step is provided in the remainder of the chapter.

Chapter 5. THE SELECTION PROCESS -- EVALUATION PRIOR TO REGULAR USE

Since the purchase of materials and equipment which support instruction is common to all three types of sub-projects, evaluation associated with the selection of materials and equipment prior to purchase is appropriate to all of them. This is a particularly crucial evaluation phase and the suggestions provided in chapter 5 along with the assistance that can be provided by the school librarian who serves on the IV-B planning committee can help to ensure that this phase is effectively carried out. The content of the remainder of the Handbook presumes that an effective selection process has been implemented, and focuses on the design of evaluations that can determine whether or not the objectives of IV-B sub-projects have been achieved.

Chapters 6-8. The Typology of Sub-Projects presented in Chapter 3 will provide readers with a guide to those portions of Chapters 6 through 8 which are appropriate to their evaluations.

CONTENTS

	Page
Foreword	iii
How to use this Handbook	v
PART I AN INTRODUCTION TO ESEA TITLE IV-B AND BASIC EVALUATION PRINCIPLES	
Chapter 1 THE CONTEXT OF TITLE IV-B EVALUATION	1
Title IV-B Legislation	1
The Concept of 'Sub-Projects' in New York State's Title IV-B Program	1
Title IV-B Evaluation Requirements	2
Constraints Affecting Title IV-B Evaluations	2
Chapter 2 OBJECTIVES AND EVALUATION	3
Why Evaluation is Important	3
How Evaluation Depends on Objectives	3
A Model For Describing Educational Programs	3
Objectives In Title IV-B Sub-Projects	5
- Performance Objectives	5
- Instrumental Objectives	5
How the Objectives Approach Can Be Used To Evaluate Materials and Equipment	7
Chapter 3 A TYPOLOGY OF TITLE IV-B SUB-PROJECTS	9
Chapter 4 A SEVEN-STEP MODEL FOR PLANNING EVALUATIONS OF TITLE IV-B SUB-PROJECTS	13

Page

PART II METHODS FOR EVALUATING ESEA TITLE IV-B SUB-PROJECTS

Chapter 5 THE SELECTION PROCESS -- EVALUATION PRIOR TO REGULAR USE 31

Chapter 6 EVALUATING INPUTS INTO TITLE IV-B SUB-PROJECTS -- MATERIALS AND EQUIPMENT 37
Records of Circulation for Educational Materials 38
Evaluating Equipment in Educational Settings 41

Chapter 7 EVALUATING THE TITLE IV-B PROCESS -- ACHIEVEMENT OF INSTRUMENTAL OBJECTIVES 45
Assessing Participant Attitudes Towards Educational Programs 46
- Numerical Rating Scales 47
- The Semantic Differential 50
Validity and Reliability of Attitude Measures 53

Observational Procedures 55

Chapter 8 EVALUATING TITLE IV-B PROGRAM OUTCOMES -- ACHIEVEMENT OF PERFORMANCE OBJECTIVES 65
Evaluating Affective Growth 66
Evaluating Cognitive Achievement 68
Using Observation Procedures to Evaluate Changes In Student Behavior 71

APPENDIX: SUGGESTIONS FOR SELECTING FROM COMMERICALLY PRODUCED INSTRUMENTS 77

SOME FINAL SUGGESTIONS 81

GENERAL BIBLIOGRAPHY ON EDUCATIONAL EVALUATION 82

Part I An Introduction to ESEA Title IV-B and Basic Evaluation Principles

CHAPTER 1
THE CONTEXT OF TITLE IV-B EVALUATION

Title IV-B Legislation

P.L. 93-380, the Education Amendments of 1974, created Title IV of the Elementary and Secondary Education Act (ESEA) to consolidate seven categorical programs from the Elementary and Secondary Education Act of 1965 into two authorizations:

Part B: Libraries and Learning Resources

Part C: Educational Innovation and Support

Part A of the Title includes the authorization, allotments, and common program requirements for the two program purposes.

Under P.L. 93-380 Part B, grants have been made to local education agencies for the following purposes:

1. The acquistion of school library resources, textbooks, and other printed and published instructional materials.
2. The acquistion of instructional equipment and materials suitable for use in providing education in academic subjects; and, minor remodeling of laboratory or other space used for such equipment and materials.
3. A program of testing students; programs of counseling and guidance services; programs, projects, and leadership activities designed to expand and strengthen counseling, and guidance services.

As a result of the ESEA Amendments of 1978, P.L. 95-561, after September 30, 1980 the Title IV-B program will no longer include provisions for programs of testing, guidance and counseling or for minor remodeling; therefore, methods for evaluating these aspects of educational programs do not appear in the present manual. The primary focus of this Handbook will be on methods appropriate for judging the effects of a program providing funds for the acquisition of library resources and instructional materials and equipment.

The term "materials" is used to refer to books, periodicals, learning kits, sound recordings, filmstrips, etc. The term "equipment" refers to educational "hardware," such as videotape machines, phonographs, slide projectors, television sets, etc. The terms "resources" and "educational products" refer to both. A further specification of eligible and ineligible purchases under these categories can be found in the 1979-80 ESEA Title IV-B Planning Guide which is available from the Bureau of School Libraries.

The Concept of "Sub-Projects" in New York State's Title IV-B Program

In New York State the Title IV State Advisory Council and the State Education Department require that the acquisition of materials and equipment be planned in a systematic manner that will contribute to local educational program improvement and be organized to respond to carefully specified needs. One of the most important elements of the policy to direct the expenditure of IV-B funds into the support of coherent program purposes has been the concept of the sub-project. The sub-project is defined as

a set of related activities directed towards the same goal, for instance, achievement of basic competencies, the attainment of library media skills, or the implementation of a career education program. School districts are required to organize their planning around one or more sub-projects and to prepare applications which include:

1. A statement of needs based on a systematic assessment of local educational conditions
2. A statement of objectives towards which the sub-project is directed
3. A statement of activities that will be conducted to help achieve these objectives, and
4. An evaluation plan for assessing the conduct and impact of the sub-project.

The preparation of applications in terms of sub-projects has resulted in an appreciable improvement in the coherence and direction of plans submitted by school districts.

Title IV-B Evaluation Requirements

Title IV-B is similar to other federally funded programs in that it has an evaluation requirement. When a school district accepts a IV-B allocation, it assumes the responsibility of assessing the extent of educational progress associated with those funds. Title IV-B does not, however, impose a particular kind of evaluation. Evaluation in Part B is governed by the assumption that districts and nonpublic schools will carry out evaluations that are appropriate to the activities which they have undertaken. Thus, Title IV-B evaluations can both meet Federal and State requirements and provide schools with timely data that can be useful for improving local educational programs. This Handbook presents principles and methods of evaluation which are relevant both to meeting the requirement to evaluate, and to carrying out evaluations that meet local information needs.

Constraints Affecting Title IV-B Evaluations

The characteristics of IV-B sub-projects impose particular requirements and constraints which make the effective evaluation of these sub-projects different from those associated with most other federally funded educational programs. In contrast to other ESEA programs, the funds available for IV-B are quite small and no funds are provided under Title IV-B for personnel or for evaluation. The size of the allocations are often too small for the funds to be considered, in and of themselves, the basis of an educational "treatment" for which dramatic outcomes can be expected. The further partitioning of these funds within the school district makes attempts to assess the impact of these funds even more difficult. These constraints have been kept in mind and a conscious effort has been made to focus on techniques and procedures which can reduce time and effort required to conduct an evaluation.

CHAPTER 2
OBJECTIVES AND EVALUATION

Why Evaluation Is Important

The purpose of educational evaluation is to provide information that can help educators make better decisions concerning the allocation of personnel and resources in order to achieve program goals.

A number of respondents to the survey of Title IV-B evaluation methods commented that it would be impossible to evaluate their sub-projects, and then went on to assert that the sub-project had been responsible for increased interest, participation or achievement in some aspect of their educational program. On what evidence were these judgements based? Evaluation is simply the process of obtaining information which may serve as evidence for just this type of assertion. This evidence can then serve as the basis for making decisions about when and how to implement, expand, improve, reduce or eliminate existing program facets in order to better achieve program goals.

How Evaluation Depends on Objectives

The type of evaluation emphasized in this Handbook involves:

1) Setting program expectations
2) Assessing the extent to which these expectations have been met by comparing them with actual outcomes, and then
3) Using the information obtained as a basis for future planning and implementation

To determine if expectations are being met it is necessary first to make useful explicit statements of objectives and then to devise means for assessing success or failure to achieve them.

A Model for Describing Educational Programs

The planning and evaluation aspects of an educational program can be viewed in terms of a three-facet model as in Figure 1.

Figure 1. Three-facet model of an educational program

The goal of educational programs is the achievement of student-related OUTCOMES -- what students will learn to do as a result of participation in the program. In Title IV-B, intended student-related outcomes are referred to as performance objectives. In order to achieve these outcomes various activities are conducted such as instruction and support services. These activities constitute the PROCESS facet of educational programs. Plans to implement such activities as part of IV-B sub-projects are referred to as instrumental objectives. Finally there are the human and material resources which make up the program. These can be thought of as the INPUT of educational programs. In Title IV-B the inputs of interest are the materials and equipment ordered with Title IV-B funds. Program INPUTS, PROCESSES and OUTCOMES represent the three program aspects that should be taken into consideration when planning and evaluating IV-B sub-projects.

The three facets of the model can be displayed as in Figure 2.

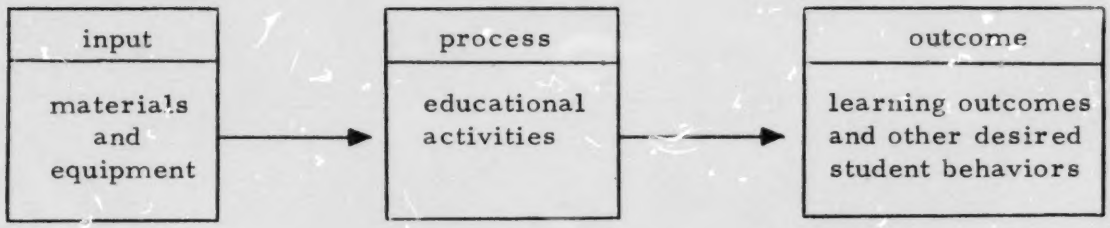


Figure 2. The three aspects of IV-B Sub-Projects that can be evaluated

This model is particularly useful in that it is congruent with the concept of performance objectives and instrumental objectives as they are used in New York State's ESEA Title IV-B Program. The nature of these two types of objectives, which play a critical role in the evaluation of IV-B sub-projects, will be discussed next.

Objectives in Title IV-B Sub-Projects

Formal statements of educational program expectations or objectives can be of two broad types. The first type of objective represents expectations for positive changes in students. These include academic achievement, skills, attitudes, self-concept and social behavior. Such expectations are called performance objectives. The second type of objective indicates expectations for improvements in the program itself including curriculum, instruction, supervision, personnel skills and resources. These expectations for conditions that will facilitate the learning process are called instrumental objectives. Instrumental objectives are expectations for what is to be done in an educational program and performance objectives are expectations for desired changes in students which are intended to result from the program. It is presumed that the achievement of instrumental objectives will lead in some evident or demonstrable way to the achievement of performance objectives. Instrumental and performance objectives are expectations for educational processes and outcomes respectively.

Performance Objectives

A wide range of possible expectations for desired changes in student abilities and behaviors are appropriate within the context of Title IV-B sub-projects. Performance objectives are generally considered to fit into one of four categories:

1. Cognitive Objectives - have to do with knowledge and intellectual skills, such as:

   - the ability to make accurate estimates of basic metric measures
   - the ability to use the catalog in searching for books on a given topic
   - the ability to write a well-structured informative paragraph

2. Affective Objectives - refer to feelings and emotions of learners, such as:

   - appreciation for classical music
   - positive attitude towards mathematics
   - commitment to self improvement

3. Psychomotor Objectives - refer to the muscular skills required for various actions, such as:

   - the ability to catch and accurately throw a rubber ball at a distance of 20 feet
   - the ability to touch type at a speed of 40 wpm
   - the ability to cut a straight line in preparing pictures for matting

4. Desired Social Behaviors. Examples of desirable social behaviors which might become the basis of performance objectives are:

   - acting with consideration for the needs of school personnel and fellow students
   - participating in communal projects
   - taking responsibility for the operation of class activities or the care of school resources

Instrumental Objectives

Instrumental objectives refer to desired conditions in educational programs which are expected to facilitate the learning process and thereby the achievement of performance objectives. The following examples are typical of the types of changes in program conditions that might serve as the basis for instrumental objectives:

- preparation of a life-skills curriculum
- setting up of a computer system so that students can have access to career education information
- conducting a series of workshops to train staff in the implementation of social studies materials.

How the Objectives Approach Can Be Used to Evaluate Materials and Equipment

The previous section on How Evaluation Depends On Objectives indicates that evaluation can be conducted by assessing discrepancies between expectations and outcomes for both instrumental and performance objectives. This method can also be extended to the evaluation of instructional materials and equipment, the input aspect of IV-B sub-projects. There are always expectations associated with the acqusition of educational products. Often these expectations are not evident because they are not stated explicitly as objectives. When they are stated explicitly, one usually can recognize that there are expectations concerning educational products that were implicitly shared. For example, there are a number of expectations which are held whenever educational materials (texts, reference books, learning kits, etc.) are used. Program personnel inevitably expect that such materials will be:

- appropriate to the curriculum which they are intended to implement
- appropriate to the level of abilities of learners
- compatible with local instructional and administrative procedures
- appealing to learners.

In the case of educational equipment; e.g., motion picture and slide projectors, cassette recorders, general expectations are that the articles of equipment will:

- be useful for the purposes intended
- be reasonably easy to use
- be durable.

When these expectations are made explicit, an evaluation plan can be developed to assess the extent to which they were met. The objectives approach to evaluation can then be used to evaluate educational products themselves as well as their role in the achievement of instrumental and performance objectives. Some suggestions for using this approach to evaluate educational products appear on pages 37-44.

CHAPTER 3
A TYPOLOGY OF ESEA IV-B SUB-PROJECTS

A wide range of educational activities could serve as the basis for a IV-B sub-project. A study of IV-B applications and responses to the survey of IV-B evaluation methods suggests, however, that sub-projects may reasonably be classified into three types. The importance of the distinction between these types is that each type may call for different evaluation procedures. Users of this Handbook are encouraged to first classify their own project into one of the three types and then, to refer to those sections of the handbook which discuss the type of evaluation they will need to conduct.

Type I

In these sub-projects there is reason to believe that the impact of sub-project activities on the achievement of performance objectives can be clearly demonstrated. Such a cause and effect relationship can be demonstrated only when the following conditions hold:

- The performance objectives refer to a set of highly specific skills, items of knowledge or attitudes.
- The objectives are attainable within the period of time given for the sub-project and its evaluation.
- There is a clearly visible relationship between the materials and equipment purchased with Title IV-B funds and the sub-project's performance objectives.
- The materials and equipment purchased with IV-B funds represent a major aspect of the instruction students will receive relating to the achievement of these performance objectives.
- The use of materials and equipment, and the amounts of instruction provided are concentrated enough so that the impact of IV-B resources can be considered sufficient to achieve the performance objectives; that is, the potential impact of the sub-project has not been diluted by spreading limited human and material resources thinly across an entire school district.

With a sub-project of this type it will be appropriate to evaluate the extent to which performance objectives were achieved.

Example of a Type I sub-project

A local needs assessment indicates that a substantial number of students in the high school lack skills for critical analysis of social issues. A decision is made to attack this problem by means of a IV-B sub-project. A committee composed of teachers, supervisors and the school librarian develop and carry out the following plan:

- o A person is selected to be responsible for the conduct and evaluation of the sub-project.
- o Objectives for student performance in critical analysis skills are prepared and approved. These are designed to be specific and to the best judgement of the committee, achievable within a half school year semester.
- o a program of instruction is outlined and designed
- o Sources such as trade journals, selection guides, bibliographies and brochures are examined for potentially useful instructional materials to support this sub-project.

- o If appropriate, samples are obtained from vendors for committee review and analysis.

- o A set of materials is selected and ordered.

- o A decision is made to limit the sub-project to a few classes deferring the decision to extend instruction to a larger group of students, pending the evaluation. The major evaluation question in this case would be "Did the sub-project enable students to achieve the performance objectives?"

- o Three committee members are asked to construct an evaluation instrument for assessing achievment of sub-project objectives.

- o An essay examination is constructed to assess these objectives and procedures for administering and reliably scoring the examination are developed.

- o Two forms of the examination are prepared, one to serve as a pretest and the other as a posttest.

- o Pretests are administered to the two classes which will participate in the sub-project as well as to two nonparticipating classes that are composed of similar groups of students. This information is intended to provide a basis for making judgments about the initial skill levels of participating and nonparticipating students before they have taken part in the sub-project. The administration of the posttest to both groups of students will then permit comparisons to be made between the two groups in terms of the achievement of the specified objectives.

  Chapter 4 A Seven-Step Model for Planning Evaluation of IV-B Sub-Projects and Chapter 8 Evaluating IV-B Program Outcomes -- Achievement of Performance Objectives provide some suggestions and sources that will assist in the planning of this type of evaluation.

Type II

While these sub-projects are also designed to contribute to the achievement of sets of performance objectives their nature is such that the contribution would be difficult or impossible to measure. This would be the case if the following conditions exist:

1) A number of factors other than Title IV-B sub-project activities and resources are major contributors towards the achievment of the set of performance objectives. Such factors might include the regular instructional program and other media resources used in instruction.
2) The set of performance objectives is too broad to reasonably expect an impact from the relatively limited IV-B sub-project. For instance, the range of skills needed to perform well on tests such as PEP, CAT, SAT, MAT, etc., are too broad to expect the implementation of a IV-B sub-project to result in dramatic improvements in such test scores for large numbers of students.

Example of a Type II sub-project:

A special remedial program is organized to improve the reading-related skills of a group of students and to supplement this program, a group of teachers, supervisors, and librarians meet to organize a sub-project which will select appropriate high-interest, low-difficulty reading materials and to identify ways to encourage students to make the most use of them. In this example IV-B resources are being used as a component of an

overall effort to increase reading achievement. Since IV-B resources are only one part of the overall effort, it is not appropriate in this situation to attempt to evaluate the sub-project in terms of an expected increase in student achievement in reading.

A more appropriate way to evaluate a Type II sub-project would be to turn to the instrumental objectives of the sub-project and attempt to answer evaluation questions such as the following:

1. Were the materials at the appropriate level of difficulty?
2. Did students find the materials interesting?
3. To what extent did students make use of the IV-B materials?
4. Were the various parts of the sub-project implemented as planned?
5. Did students read more and make greater use of the library?

In cases such as the above where it is unreasonable or impractical to attempt to evaluate the sub-project in terms of the achievement of performance objectives, the process aspects of the sub-project (i.e., the achievement of instrumental objectives) can be evaluated as well as the quality, appropriateness, appeal and extent of use of the materials themselves. The following sections of the Handbook provide suggestions and procedures for conducting such evaluations:

Chapter 6 Evaluating Inputs Into IV-B Sub-Projects -- Materials and Equipment, and
Chapter 7 Evaluating The IV-B Process -- Achievement of Instrumental Objectives

Type III

These sub-projects meet general instructional or learning needs rather than contribute to the achievement of a fixed set of performance objectives. In sub-projects of this kind it is not practical or appropriate to try to associate student outcomes with IV-B resources. This would be the case when a wide range of supplies is ordered for a science laboratory, reference books are ordered for a library or a piece of audiovisual equipment is required for use across a variety of subject areas. The type of school library or instructional resources acquired in this kind of project might include items as a motion picture projector, a set of scales or a microscope for the science lab, a video-tape machine and camera, or a set of encyclopedias for the library.

Example of a Type III sub-project:

School librarians received numerous inquiries from students looking for more detailed information on a variety of topics than is available from existing sets of encyclopedias. The librarian developed a survey form designed to elicit statements from students who used the library for research purposes concerning their reference needs and their perceptions of the adequacy of available sources. Based on the need defined through this survey, IV-B funds were used to acquire a more advanced set of encyclopedias as well as some important reference works related to subjects about which students made frequent requests.

Evaluation appropriate for this kind of project could include maintaining a record of the frequency with which the materials are used, an indication of the extent to which users found the material appropriate to their abilities and the degree to which it satisfied their needs for information about a particular topic.

Chapter 6, Evaluating Inputs Into Title IV-B Sub-Projects - Materials and Equipment, contains some ways to improve "frequency of use" information which are based on suggestions provided by respondents to the survey of IV-B evaluation methods. Another way of evaluating such sub-projects is to survey the attitudes of teachers and students towards use of these resources in the local educational context. Methods for studying such attitudes are provides in Chapter 7 - Evaluating The IV-B Process -- Achievement of Instrumental Objectives.

CHAPTER 4

A SEVEN-STEP MODEL FOR PLANNING EVALUATIONS OF TITLE IV-B SUB-PROJECTS

The following sequence of steps is suggested as a model for planning a Title IV-B sub-project evaluation. The sequence outline is followed by discussions which explain in greater detail what is involved in implementing each of the steps. The unique characteristics of individual sub-projects may necessitate different orderings of steps and varying degrees of attention to each, but in general, the sequence should be appropriate for planning evaluations for all three of the sub-project types described in Chapter 3.

The ordering of the steps reflects the fact that in planning a process as opposed to implementing that process one must often consider "last things first." That is, one must consider the nature of the desired end product and the steps leading to this end before initial and intermediate steps can be adequately conceptualized.

1. Set objectives in a way that will facilitate evaluation
   a. make objectives specific
   b. set levels of expectation for achievement.
2. Determine information needs. Those objectives which are most closely related to the aspects of your program that require data collection are the ones on which evaluation efforts should be focused so that the information collected will be useful for decision making.
3. Consider how evaluation information will be utilized.
   a. suit evaluation information to the needs of varying audiences
   b. make provisions to meet the timelines of formative and summative information needs.
4. Determine how desired information will be analyzed and presented.
5. Develop or select an instrument to obtain desired information. It is advisable to consider how the information will be analyzed and presented before developing an instrument. This will help to assure that the instrument can, in fact, generate the desired information.
6. Plan data collection procedures.
7. Collect base-line information and obtain measurements from a comparison group.

This entire series of steps is most profitably conducted during the planning phase of a sub-project. Then problems in the evaluation plan such as unrealistic objectives or an ineffective instrument, can be worked out while the sub-project is being carried out. Putting these efforts off until the time when data must be collected decreases considerably the chances of obtaining useful information.

The remainder of the present chapter provides more extensive discussions for each of the seven steps for those who wish more explicit directions. Examples illustrating steps four through seven appear in the following three chapters:

Chapter 6 Evaluating inputs Into IV-B Sub-Projects -- Materials and Equipment

Chapter 7 Evaluating the IV-B Process -- Achievement of Instrumental Objectives

and

Chapter 8 Evaluating IV-B Program Outcomes -- Achievement of Performance Objectives

STEP 1. Set objectives in a way that will facilitate evaluation

a. make objectives specific
b. set levels of expectation for achievement.

The following discussion explains why it is desirable, from the point of view of evaluation, to have objectives which are specific and which include levels of expectation for achievement. The discussion concludes with a presentation of the official format for writing objectives in Title IV-B applications.

Setting reasonable expectation for growth and change assures that clear evidence will be available when these expectations are met. For example, when a set of objectives similar to the one shown below is incorporated into an organized unit of instruction and supplemented by instructional materials appropriate to meeting such an objective then it is reasonable to expect not only that the objectives can be achieved but also that means can be devised to provide evidence of their achievement.

- o By June 1979, a (specified) proportion of students who participate in the basic competency sub-project will be able to determine sale prices when provided with the list price and integral discount rate.

Making explicit statements about levels of expectation for sub-project objectives also makes it easier to determine when the objectives are met. The following objectives specify explicit standards:

- o students will be able to answer correctly at least 75% of the questions in the punctuation and usage portions of a school-wide language examination

- o by February 1, 1979 the utilization of taped programs in (specified) instructional areas will have increased 15% over the previous year.

When standards or levels of expectation are stated explicitly in this fashion, it will be quite clear whether or not objectives have been attained and one can usually avoid dependence on inferential statistics (see STEP 4) as a means for quantifying one's level of confidence in their attainment. Base line performance measures and needs assessment results can be valuable sources of information for setting realistic standards of expectation.

Writing Performance Objectives for Title IV-B Sub-Projects

In the 1979-80 ESEA Title IV-B Planning Guide, a format is presented for stating instrumental and performance objectives in Title IV-B applications. The format specifies that each performance objective should contain the following:

- o A description of what learners will be able to do as a result of participating in program activities
- o A statement of the criterion that will be used to determine how successfully the objective has been achieved
- o A proposed date for achievement of the objective
- o A statement of the percent (%) of learners who are expected to achieve the objective by the specified time.

The following performance objective would be appropriate for a district which has used its IV-B allocation to purchase two computer terminals and related instructional

materials for the purpose of implementing a sub-project to teach students fundamental programming skills:

> By June 30, 1980 all of the students enrolled in the Computer Lab Program will be able to write and successfully execute a computer program based on a set of specifications and data provided by the instructor.

In another example of a high school science elective project a camera was purchased with Title IV-B funds and used to supplement instruction in astronomy, environmental ecology and advanced biology as well as in the art of photography. The performance objective for this Title IV-B sub-project was stated as follows:

> By June 1979, students who have participated in the science elective program will have produced photographic essays, which meet a set of criteria specified by the instructor, indicating that they recognize various astronomical phenomena and selected elements of the environment and ecology of the Dunham's Bay Wetlands.

Writing Instrumental Objectives for Title IV-B Sub-Projects

The 1979-80 ESEA Title IV-B Planning Guide specified a format for writing instrumental objectives which requires the following:

- A description of the changes to be brought about in local district conditions
- A proposed date for achievement of the objective
- A statement of the method that will be used to determine the extent to which the desired changes occurred
- A statement relating the instrumental objective to a performance objective or to a general program goal.

An example of an acceptable instrumental objective is:

> By December of 1980 all science teachers in the middle school will have completed training in the use of the "Shrinking Universe" learning program and will have registered their reactions to the quality of the materials on a locally prepared rating scale.
>
> The achievement of this instrumental objective is intended to support the achievement of program goals related to knowledge of developments in human technology.

When objectives are stated in the above manner it is a relatively straightforward process to design a means to assess whether or not they have been achieved.

STEP 2. Determine information needs.

The second step of this process is to determine which objectives are most closely related to aspects of your program that require data collection. These are the objectives on which evaluation efforts should be focused so that the information collected will be useful for decision-making.

The typology pesented in chapter 3 provides a guideline for choosing which aspect of a sub-project to evaluate based on characteristic properties of the sub-project. Another important consideration in deciding what to evaluate is the nature of local information needs.

Since resources for evaluating Title IV-B sub-projects are limited it will usually be desirable, and often necessary, to focus evaluation efforts on a particular aspect of the program. To decide which aspects to evaluate one should first consider what use will be made of the evaluation and what form the information should take to be maximally useful.

Determining Decision-Making Needs and Posing Evaluation Questions

As an example consider a sub-project which involves the acquisition of elementary school instructional materials in mathematics. The needs assessment conducted as part of the IV-B planning process in a local school district indicates a general deficiency in students' abilities to make use of the principles of multiplication in problem solving applications, particularly in problems involving the surface areas of polygons. The planning committee charts a sequence of desired skills which they feel will remedy the deficiency and plans an instructional program in which new materials relevant to problem solving in mathematics can be incorporated into the already functioning learning lab. In addition, a series of workshops is organized to introduce teachers to the sub-project aims, use of the new instructional materials, and procedures for making most efficient use of the learning lab. The "inputs" into this sub-project would be the instructional materials, the "process" would consist of instructional activities and the "outcomes" would be the skills which students are expected to achieve as a result of sub-project implementation.

District personnel might have any of the following decisions to make, related to the above sub-project:

- Whether or not to use the instructional materials again
- What level of students to use the materials with
- Which schools to use the materials in
- Whether or not to provide more extensive training for teachers in using the materials and equipment.

These can be expressed as evaluation questions which provide a more specific indication of what type of evaluation information must be collected:

- Did students meet the specified performance objectives for the sub-projects?
- Did they do so according to tests that accompany the kits, according to school or district-wide criterion or both?
- Did teachers follow the prescribed procedures for using the materials? How did they feel about these procedures and the materials themselves?

When decision-making needs and evaluation questions are stated explicitly in this fashion they can be ranked in terms of which information is most crucial and unecessary evaluation efforts can be avoided. For example, in a district where materials such as the ones described above have been used successfully for a number of years in different schools and are now being applied to a new school for the first time, an input evaluation might be considered unnecessary and evaluation of teacher workshops, program implementation and the achievement of cognitive objectives emphasized.

CARR V VOC III M-E VOCABULARY GRADE 10 5/79
FORM # 1 STUDENT PROFILE

QUESTION #'S --->	0000000001 1111111112 2222222223 333			
	1234567890 1234567890 1234567890 123	-	%	
ANSWERS -------->	DEAEDBBCDA CEBEDCBDAA DCAABBEDDE CCC			
student names				
	---------- ---B-A-C-- BA-DC---/- --D	24	73	
	---------- -C--------C ---D----E- BB-	27	82	
	---------- -------B-- ---DA---EA ---	28	85	
	---B------ ---------- A-CD------ ---	29	88	
	-D---E---- -------B-- B---D--BE- --E	25	76	
	-----C---- --------C- -----DB-ED E--	26	79	
	---------- -----B--C- --------E- E--	29	88	
	---------- --A-----C- -------EA- ---	29	88	
	---------- --A----C-- --------E- ---	30	91	
	-----C-A-- -----BC--- B--D-----EA D--	24	73	
	---------- --------C- -------A-- ---	31	94	
	---------- ---B----C- A-------EA ---	28	85	
	---------- ---------- --------ED ---	31	94	
	---------- -------B-- A--D---BA- ---	28	85	
	-----D---- --A----BC- --------E- E--	27	82	
	---------- -C---B--C- A----C--A- ---	27	82	
	---------- -------BC- ---------D ---	30	91	
	---------- -------B-- -----C-E-- ---	30	91	
	---------- ---------- ---------- ---	33	100	
	---------- -----A---C ---D------ ---	30	91	
	---------- ---------- --------BA ---	31	94	
	---------- --A--B--C-- ---B---EB- E--	26	79	
	---------- ---------- --------E- ---	32	97	
	---------- -----A-B-- --BD----C- E--	27	82	
OBJECTIVE ------->	0000000001 1111111112 2222222223 333	↑ raw score	↑ percentage	
	1234567890 1234567890 1234567890 123			

KEY -: the student answered the question correctly
/: the student did not try this question and left it blank
letter: the student missed this question and the letter response indicated his response

Figure 3a Analysis of a Multiple Choice Test for a Class (Courtesy Maine-Endwell School District

STEP 3. Consider how evaluation information will be utilized

Suit evaluation information to the needs of varying audiences

Often there will be several different audiences awaiting the results of an evaluation. Very different reports on the findings of a program evaluation might have to be prepared to meet the information needs and concerns of groups as potentially diverse as teachers, school administrators, librarians, students and school board members. A report which is not "on target" for a person's or a group's concerns and their level of sophistication with regard to a particular issue could easily be neglected.

Make provisions to meet the timelines of formative and summative information needs.

Timelines should be developed in concert with planning and decision making needs. The question of timeliness is closely connected to whether formative or summative evaluation is being conducted. A formative evaluation is the assessment of a program in progress for the purpose of making decisions about how to improve its operation. A summative evaluation is one which is conducted at the end of an educational program to determine the extent of its success and to isolate which factors may have been responsible for success or failure.

The computer printouts displayed in Figures 3a and 3b are an interesting example of formative evaluation drawn from a Title IV-B sub-project. The performance objectives for the sub-project were focused on vocabulary development for grades 7-11. Vocabulary texts were ordered, a system of lessons planned and tests generated at each level. The tests were scored and analyzed by computer and teachers received printouts of results for classes and individual students two days after the test was taken. This meant that evaluation results were provided to teachers soon enough to help them plan short- and long-term instruction for both individuals and classes.

Figure 3a shows a class analysis of a multiple choice vocabulary test allows the instructor to see at a glance:

- o Which items tended to be answered correctly or incorrectly
- o Patterns of incorrect answers which could suggest needed insructional emphasis or possibly test item improvement
- o What objectives the various test items were related to
- o The relative status of students.

Figure 3b shows an individual test analysis for a student which comes in the form of a sticker that can be attached to the student's records. It too tells which items the students answered incorrectly, providing an information base for individualized planning.

STICKER

```
000000000111111111122222
123456789012345678901234 5
BDCCABDEDDCACBCDBACBABBAB
----------------------C-D
BLINK  J: TALLY 23 , MARK= 92 ,  #
```

--answers (number and correct ans.)

--student responses

Figure 3b. Analysis of a Multiple Choice Vocabulary Test for One Student (Courtesy of Maine-Endwell School District)

Since current Title IV-B regulations permit sub-projects to extend for up to a three-year period, serious consideration should be given to the advantages of conducting formative as well as summative evaluations.

STEP 4. Determine how desired information will be analyzed and presented

During the actual conduct of an evaluation, the data analysis and presentation must necessarily folow the development of an evaluation instrument. For planning purposes, however, it is wise to carefully consider the way information will appear in its final analyzed state before developing or selecting an instrument. This is to help assure that the evaluation instrument finally chosen or developed can actually produce this type of information.

The Use of Statistical Analyses

A wide variety of techniques appropriate to the analysis of evaluation data are available in introductory texts on applied statistics such as the ones referred to at the end of this step. The statistical methods presented in these texts can be used in either a descriptive or an inferential manner.

Descriptive statistics involve systematic procedures for the organization, summary, and analysis of quantitative information. The analysis of sets of scores in terms of their means, medians, percentiles, percentile ranks, etc., represents an extremely useful tool for summarizing and interpreting test scores. An evaluator who wishes to be able to interpret results from standardized tests or to conduct a systematic examination of scores derived from local measures will find the basic principles and terminology of descriptive statistics invaluable.

The types of measures used in education (classroom and standardized test scores, observational ratings, attitude ratings, etc.) tend to be somewhat undependable, often containing substantial errors of measurement of various types. The generalizations that we wish to make about scores which represent group and individual characteristics are most appropriately viewed as based on estimated rather than true values. Inferential statistics are valued in educational research because they permit numerical estimates of the levels of confidence that we can have in our generalizations concerning these scores. These confidence estimates are based on consideration of the extent to which findings might be attributable to sampling and measurement errors and other uncontrolled factors rather than to the factor of interest; that is, contributions of IV-B sub-projects.

Because of the complex interrelationships between factors which may influence the course of educational programs and the nature of the inferential requirements and types of measures used, educational research often requires complex research designs and analyses involving the use of inferential statistics. The evaluator of a Title IV-B sub-project usually does not need to make generalizations beyond the group being studied. When this is the case, and the growth or effect measured is dramatic enough so that it cannot reasonably be attributed to chance or to errors of measurement, it is not necessary to employ inferential statistical procedures and the highly structured research designs needed to meet the assumptions underlying their use. The seventh step of this model in particular is intended to help obviate, as much as possible, the need for complex analytic procedures by simplifying the scope and intention of IV-B sub-project evaluations. Specific methods for analyzing and displaying evaluation data are presented in chapters 6 through 8 of this Handbook. For those who wish to become familiar with methods of statistical analysis used in contemporary educational research the following bibliography may be useful.

Chapter 4 - A Seven-Step Model for Planning Evaluations

District Evaluator's Handbook of Selected Evaluation Procedures. The State Education Department. Bureau of School and Categorical Programs Evaluation. Albany, New York, 1976.

Ferguson, G.A. Statistical Analysis in Psychology and Education. (4th ed.) New York: McGraw-Hill, 1976.

Fitz-Gibbon, C.T., and Morris, L.L. How to Calculate Statistics. Beverly Hills: Sage Publications, 1978.

Glass, Gene V., & Stanley, J.C. Statistical Methods in Education and Psychology. Englewood Cliffs, N.J.: Prentice-Hall Inc., 1970.

Graphs and Statistics: A Resource Handbook. The State Education Department. Bureau of General Education Curriculum Development. Albany, New York, 1977.

Roscoe, J.T. Fundamental Reserach Statistics for the Behavioral Sciences. (2nd. ed.) New York: Holt Reinhart and Winston, 1975.

STEP 5. Develop or Select an Instrument to Obtain Desired Information Needs

The term "instrument" refers to the device used to collect evaluation information. The specific instrument used may be an achievment test, an attitude survey, an observation schedule, a circulation record form or any of a number of other data collection devices. It is advisable to consider how the information will be analyzed and presented before developing an instrument. This will help to assure that the instrument can, in fact, generate the desired information.

There are two ways to collect information about educational programs. One way is to elicit responses of participants and other interested parties by means of tests, surveys, questionnaires, interviews, etc. The other way is by means of nonreactive measures, or as they are sometimes called, nonobtrusive measures. Nonreactive measures include procedures for observing and recording various aspects of educational processes, and on-going programs (circulation statistics, records of frequency-of-use of equipment, records of attendance of participants at various functions, etc.). In general there are two desirable properties of instruments used to collect information. These are called validity and reliability.

Validity and Reliability of Evaluation Instruments

When an instrument provides the information that it was intended to, it is considered valid. For instance, the validity of the information from a survey designed to record student attitudes towards a film series on electricity would depend on how well it captured those attitudes. It is usually quite difficult to assess the validity of an instrument. Often another independent measure of the same attribute (in this case student attitudes) would have to be obtained and an attempt made to see how closely the two measures agree.

The validity of the information derived from an instrument is relative to the purposes for which it is used. For instance, a method could be developed for recording the proportions of time that teachers spend implementing various components of a structured learning package. This instrument would be valid for the descriptive purposes intended but would not be a valid one if it alone was used for the purpose of evaluating the teacher's success in meeting curricular objectives. Validity thus has to do with the appropriateness of an instrument to the purposes for which it is being used.

An aspect of validity which deserves greater attention is "content" or "curriculum" validity. This type of validity has to do with the appropriateness of a test to a given set of learning objectives such as a given curriculum unit. Classroom teachers are often among the most valuable sources of judgement concerning this type of validity.

Reliability has to do with the dependability of the measures obtained with an assessment instrument. A reliable instrument is one which provides information which is relatively free of fluctuations due to irrelevant factors.

The ratings given by three independent observors during the same thirty minute period, on an instrument designed to assess how much class time was devoted to the achievement of a given performance objective, would be expected to be in agreement. To the extent to which they were in agreement, the information obtained from this instrument could be considered reliable. If the ratings are substantially influenced by temporary conditions such as mood, level of attention or effort of the raters, or inadequacies in the rating procedure which permit varying interpretations of what is being observed, then the information obtained would tend to be unreliable and the ratings given by different observers would tend to disagree. The extensive procedures used in the administration and scoring of standardized tests serve to maintain the relability of these instruments by controlling for the effects of irrelevant factors on test performance.

It is desirable to make some effort to assess the validity and reliability of an evaluation instrument so that judgements and conclusions based on data can be qualified by a recognition of the instrument's potential limitations. Techniques for quantifying estimates of validity and reliability can be found in the introductory textbooks on educational testing, measurement and evaluation cited below.

Anastasi, A. Psychological Testing. (4th. ed.) New York: MacMillan, 1976.

Brown, F.G. Principles of Educational and Psychological Testing. Hinsdale, Illinois: Dryden Press, 1970.

Cronbach, L.J. Essentials of Psychological Testing. (3rd ed.) New York: Harper, 1970.

Guilford, J. Psychometric Methods. (2nd ed.) New York: McGraw Hill, 1965.

Mehrens, W.A. and Lehman, J.J. Measurement and Evaluation in Education and Psychology. New York: Holt, Rinehart and Winston, 1975.

Siegel, S. Nonoparametic Statistics for the Behavioral Sciences. (2nd ed.) New York: McGraw Hill, 1974.

STEP 6. Plan Data Collection Procedures

Facilitating the Process of Data Collection

Evaluation procedures should be as nonobtrusive as possible and the burden of expended time and effort on the part of staff and students kept to a minimum. Still, some time and effort is required, so your sources of information should be approached with sensitivity, respect, and gratitude for their contribution to the evaluation effort.

Figure 4, showing cover sheets to survey forms, suggests some appealing ways to request evaluation information.

Dear Faculty,

Since time is so limited for everyone these days, I do not intend to burden you with endless questionnaires. However, it is essential that I get some information from you regarding your feelings about library service. Please complete the attached questionnaire and return it to me at your earliest convenience.

Many thanks,

Mary Emerson

MEDIA CENTER

March 27, 1978

Dear Students,

I really appreciate you and your teacher taking the time to answer a few questions for me. It is very important for me to know what you want and need in order to make the library useful to you. At any time during the year if you have ideas about what the library should have, please feel free to talk to me about it.

It is not necessary to put your name on the questionnaire, but please state your grade level.

Once again, thank you very much for your help.

Sincerely,

Mrs. Emerson
Librarian

Figure 4 Cover sheets requesting evaluation information. (courtesy of the Brunswick Central School District)

Reducing Data Collection and Analysis Efforts

The need to minimize the efforts required of respondents in data collection was voiced numerous times in response to the survey of IV-B evaluation methods. The use of structured response (e.g., true-false or multiple choice items) on evaluation instruments was the means most often suggested for reducing the burdens of data collection. The alternative to structured response formats is some type of narrative response. There are advantages to and disadvantages to both types of responses.

Advantages of a structured response format:

- Items can be designed to provide precisely the kind of information that the evaluator is prepared to interpret
- The processes of data tabulation and analysis are greatly simplified, especially if the analysis plan has been designed before the evaluation instrument is constructed
- The task of filling out the instrument is easier for respondents or observers.

Disadvantages of a structured response format:

- A valid and reliable structured response instrument may be a long time in the making, requiring continual revision and replacement of items until they meet the evaluator's requirements
- A structured response form does not include provisions for issues and points of view which may be relevant but which the researcher did not take into consideration.

Advantages of a narrative response format:

- A narrative response format has the advantage of permitting respondents the freedom to express their opinions and interpretations concerning an issue.

Disadvantages of a narrative response format:

- The time it takes to provide narrative rather than short answer responses
- Narrative responses may be filled with irrelevant commentary leaving the issue at hand unanswered
- Prolonged reading and content analysis of narrative responses may be required to extract desired information
- Extensive efforts must often be made to prepare a reliable analysis system (one that will result in reasonable agreement between different persons analyzing the same responses).

There are two types of compromises which can balance these relative advantages and disadvantages:

1. Make the request for a narrative response in itself highly structured, directing the respondents or observers to provide needed information precisely, and a systematic analysis plan is prepared to assimilate this information.
2. Invite respondents to a structured response instrument to make whatever narrative comments they feel might be appropriate.

STEP 7. Collect Baseline Information and Obtain Measures from a Comparison Group

Numerous factors can effect the conditions found in educational programs. The purpose of evaluation design is to attempt to assess which factors are responsible for success in bringing about an expected outcome or for failure to achieve that outcome. Evaluation design usually involves an attempt to assess the effects of one or more factors by arranging conditions of study or administration so that the effects of other factors are controlled.

For example, in order to assess which of several modes of instruction (e.g., group discussion vs. directed research) is most effective in leading to the achievement of the same set of objectives each of these methods might be implemented simultaneously with similar and separate groups of students. Provisions could be established to assure that each group received equal amounts of instruction in order to control for varying effects of participation. An attempt might also be made to alternate instructors from one method to another in order to control for the effects of teacher style or personality on the achievement of learning outcomes. By means of controls such as these, efforts are made to isolate, as much as possible, the role of the instructional modes, independent of other factors, in helping to bring about intended outcomes.

When a program objective implies an expected growth in some area of knowledge, skill or attitude, baseline measures of this attribute should be obtained. Baseline information in IV-B sub-projects refers to measures of personnel, program and student attributes that exist before the sub-project is instituted. Pretests given by classroom teachers are an example of baseline information. Needs assessment information that is systematically collected and analyzed can also serve this purpose. Without baseline information, judgements about growth or changes in individuals and programs cannot be made. This is another reason why evaluation should be designed in detail during the planning of a sub-project.

It is useful to have a comparison group to serve as a control factor in the evaluation. The examination of pre- and postscores on the evaluation instrument obtained from a group of students similar to sub-project participants, but who have not received the sub-project treatment can reveal much about the properties of the evaluation instrument and the chances of achieving sub-project objectives even without participation. If similar results are obtained whether or not the Title IV-B sub-project is present, one would have reason to suspect either the validity of the evaluation instrument or the claims of the sub-project.

The following sources may serve as introductions to issues of design in evaluation research:

Campbell, D.T. and Stanley, J.C. "Experimental and Quasi-Experimental Designs for Research on Teaching." In N.L. Gage (Ed.), Handbook of Research on Teaching. Chicago: Rand McNally, 1963. Reprinted as Experimental and Quasi-Experimental Design for Research. Chicago: Rand McNally, 1960.

Fitz-Gibbon, C.T., and Morris, L.L. How to Design a Program Evaluation. Beverly Hills, Sage Publications, 1978.

Kerlinger, F. Foundations of Behavioral Research. (2nd ed.) New York: Holt, Rinehart and Winston, Inc.

PART II METHODS FOR EVALUATING ESEA TITLE IV-B SUB-PROJECTS

Introduction to Part II

In Part I of this Handbook a rationale and an outline for the conduct of systematic evaluations of educational programs was presented. What follows, Part II, provides a guide to specific techniques of data collection and analysis which can be used within the context of a systematic evaluation of Title IV-B sub-projects.

It is important to recognize that similar and sometimes identical instruments can be used for disparate purposes. For example, one might collect statistics on the number of times educational resources ordered with Title IV-B funds were borrowed from a media center. This information could be one piece of information among several that would contribute towards an assessment of the quality and interest value of the materials. On the other hand, by taking a group of students participating in a Title IV-B project which is intended to spark their interest in reading, and examining the frequency with which they borrow books and other media resources, before, during, and after the program, one would have information that could contribute to understanding the impact of the program. For this reason, whenever a method is presented, mention will be made of the various aspects of Title IV-B evaluation to which it might be applied.

The principles and methods presented in this Handbook are intended to be suggestive rather than prescriptive. It is not expected that any method described in this booklet will be duplicated exactly. Readers are encouraged to focus on those techniques which are most relevant to their purpose, and to be sensitive to the need for adjusting or redesigning them to fit local requirements.

The following discussion contains suggestions for collecting information to evaluate Title IV-B sub-projects. Included are descriptions and examples of methods used by people who coordinated IV-B projects and some drawn from other sources. The range of suggestions and possible evaluation facets should permit those who are going to conduct IV-B evaluations to choose a facet and method most relevant to their decision-making needs and to introduce whatever modifications are required by the context of their educational program.

CHAPTER 5
THE SELECTION PROCESS - EVALUATION PRIOR TO REGULAR USE

It is expensive and inefficient to discover, after ordering school library and other instructional resources, that they do not meet the expectations held for them. For this reason it is important to systematically evaluate materials and equipment prior to purchase.

The following suggestions are presented to assist in the collection and efficient use of information when making selection decisions:

1. Include a Wide Range of Persons in the Selection Process

More than 61 percent of the sub-projects surveyed included both users of educational products and their supervisors in the process of examining literature and samples of products when making selection decisions. One of the most consistent observations made by respondents to the survey was that broadening the base of participation increased the quality of selection decisions.

2. Assess the Appropriateness of Materials and Equipment to Students for Whom it is Intended.

The assessment of the appropriateness of materials to learners is often done by matching suggested grade levels for materials with students' grade level scores on standardized tests. If the latter are not available a number of procedures exists for assessing student reading ability levels in terms of grade levels (Dale & Chall, 1948), (Fry, 1968), (Harris & Sipay, 1975).

Estimates of abilities expressed in terms of grade levels can vary with the degree of student interest in the reading materials as well as with the method of assessment used to determine grade levels. It is desirable to conduct as wide a range of assessment of the appropriateness of materials under consideration as is possible. When, for example, a learning package includes mastery tests it is worth examining how a sample of students from the target population perform on the initial and final exams before they have received instruction.

3. Assess the "Goodness of Fit" of Materials Under Consideration to Curricular and External Criteria

Materials designed for instructional purposes can be evaluated against two broad sets of criteria which might be termed external and internal. External criteria would be the basis for judgements concerning the soundness of subject matter presentations and for judgements concerning the extent to which the materials may be racist, sexist, or deficient with respect to some other sociocultural criteria. Internal criteria are those used to assess the appropriateness of instructional materials to the existing curriculum, instructional program and administrative procedures in a school district.

A State Education Department publication, Selection Guidelines, available from the Bureau of School Libraries, can serve as a valuable resource in setting up systematic procedures for making decisions along the lines of both internal and external criteria.

Guidelines for Instructional Materials Analysis

I. The Producer
- A. CREDENTIALS
  - 1. Reputation
  - 2. Author's education and experience
- B. CLAIMS FOR THE MATERIALS
  - 1. Target population
  - 2. Conditions of use (See II.F)
- C. DEVELOPMENTAL HISTORY OF MATERIALS
  - 1. Developmental testing
  - 2. Field-testing of final version
  - 3. Continued learner verification for revision.

II. Instructional Design
- A. GOALS AND OBJECTIVES
  - 1. Degree to which specified
  - 2. How stated or described
  - 3. Bases for selection
  - 4. Aspects of career education and/or other areas emphasized
- B. CONTENT
  - 1. Scope
    - a. Degree to which specified
    - b. Selection and arrangement
    - c. Bases for selection
    - d. Aspects of career education
    - e. Validity and accuracy of content
    - f. Fairness in terms of sex roles and socioeconomic, ethnic, regional, and national backgrounds (See Chapters 2 and 3)
  - 2. Sequence
    - a. Degree to which specified
    - b. Bases for determining
    - c. Aspects of career education and other areas related to
- C. METHODOLOGY OF LEARNING/ INSTRUCTION
  - 1. Degree to which specified
  - 2. Reading approaches identified with
  - 3. Teacher-pupil-other roles
  - 4. Organization, space, time, and grouping arrangements
  - 5. Media for presentation
  - 6. Bases for determining
- D. EVALUATION AND ASSESSMENT
  - 1. Degree to which specified and supplied
  - 2. Approaches used (e.g., norm- or criterion-referenced, etc.)
  - 3. Reading and other areas covered
  - 4. Who carries it on and when
  - 5. How recorded, used, reported
- E. DESIGN CONGRUENCIES
  - 1. Consistencies among design elements, A-D
  - 2. Congruency of each element and total design with:
    - a. Target population
    - b. Teacher characteristics
    - c. Overall organization of classroom and school program, including other subject areas
    - d. Community needs, interests
- F. CONDITIONS OF USE
  - 1. Required supplementation:
    - a. Materials
    - b. Equipment
  - 2. Personnel
    - a. Additional staff, aids, volunteers
    - b. Personnel competencies training
  - 3. School organization
    - a. Teacher loads and schedules
    - b. Time and space requirements
    - c. Budget and other resources required

III. Evidence of Effectiveness
- A. PRODUCER'S EVIDENCE (See IC above and Chapter 4)
  - 1. Opinion
  - 2. Evidence from teachers
  - 3. Evidence from students
- B. REPORTS FROM OTHER USERS
  - 1. Anecdotal
  - 2. Systematic empirical
    - a. Criterion tests
    - b. Interviews and questionnaires
    - c. Observation
  - 3. Emphasis
    - a. Effect (on learners)
    - b. Affect
    - c. Conditions of use and usefulness, or ease of use

IV. Overall Assessment and Recommendations

Figure 5

From EPIE career education s*e*t* (1975)

The 'goodness of fit' between instructional materials and program facets can be established in four broad areas:

- o goals and objectives
- o scope and sequence
- o instructional methodology
- o evaluation and assessment.

This category system applied to the evaluation of instructional materials has been evolving for a number of years (Eash, et. al. 1975) and is currently used by the Education Products Information Exchange (EPIE). Figure 5 shows an expanded outline of the system incorporating both internal and external criteria applied to the analysis of instructional materials in career education. Those features of the outline which are deemed relevant to the concerns of a local school district can be used as a basis for the design of checklists or survey forms to assess the extent to which materials meet specified criteria.

Figure 6 shows an evaluation instrument devised by a Title IV-B planning committee which provides information relating instructional materials to given objectives, stage of instruction and level of learner abilities.

Materials (Title) _____ Audience Individual

Teacher _____ Small Group

Grade _____ Class

Students Above Average
Involved:

Average

Below Average

Mixed

Usage (Circle One)

1. Introduction 2. Basic Instruction 3. Reinforcement 4. Supplementary

Circle grant objective(s) which are most appropriate according to the way you use the materials.

I. A or B
II.
III. A, B, or C
IV. A, B, or C
V. A, B, C, D, E, or F

Additional uses and/or comments _____

Figure 6
An Instrument for Evaluating Instructional Materials in Terms of Sub-Project Objectives
(Courtesy of Rush-Henrietta School District)

The sub-project in this case was aimed at improving the reading skills and motivation of underachieving students. Six objectives and a number of sub-objectives (A through F) were set towards this end. The evaluation form provides a basis for relating information on the appropriateness of the IV-B materials (and undoubtedly other materials as well) to specific objectives or sub-objectives, to particular types of students, and to stages of the instructional process. The Title IV-B coordinator for this sub-project reported that this led to the discovery of new uses of instructional materials and assisted in maximizing the use of worthwhile materials.

4. Make use of the Possible Services Available from Producers and Distributors of Educational Products.

Sales representatives often provide review copies of materials, and when available documentation of "field tests," "validation studies," or "learner verifications and revisions" that have been conducted on the products they are selling. It may also be useful to ascertain what success other schools and school districts have had in using the materials and equipment under consideration.

5. Make Use of Standard Media Selection Aids

The United States Office of Education, Office of Libraries and Learning Resources has published Aids to Media Selection For Students and Teachers, 1979, a comprehensive list of bibliographies and journals which review books, periodicals, and audiovisual materials of relevance for elementary and secondary school instructional programs. Single copies are available free from:

Office of Libraries and Learning Resources
United States Office of Education
400 Maryland Avenue, S.W.
ROB 3, Room 3125B
Washington, D.C. 20202

The nonprofit Educational Products Information Exchange (EPIE) provides a unique range of services in its role as the "Consumers Union" of educational products. EPIE publishes regular Materials and Equipment Reports. The Materials Reports provide comparative analyses of instructional materials and packaged programs in a wide variety of content areas. For instance, one report theme might be Elementary School Language Arts Programs and a number of such programs would be reviewed. The EPIE Equipment Reports present laboratory tests procedures and findings for evaluating particular pieces of equipment that are commonly used in educational settings, for example, motion picture projectors.

Twice monthly, EPIE publishes a newsletter called EPIEgram keeping consumers of educational products current on developments which may be of interest to them. It also serves as a forum through which users may relate their experiences with the quality of particular educational products.

Subjects such as the following are regularly covered in EPIEgram.

- hints on efficient use and maintenance of materials and equipment
- news of important technical developments related to educational products
- reports on patterns of malfunction in particular brands and types of equipment
- reports on surveys of the status and needs of consumers of educational products
- discussions on issues affecting educational consumers (e.g., controversies concerning the use of calculators for mathematics instruction, implications of copyright laws, controversies concerning the content of textbooks).

EPIE also conducts workshops to train local personnel in the selection and evaluation of educational materials and equipment. A number of BOCES subscribe to EPIE services and make them available to their component school districts.

6. Consult with your Library Media Specialist

The school library media specialist is an expert in the selection of resources and has access to wide-range bibliographies and reviewing media. Procedures should be established to provide continuing consultation among teachers, administrators and school library media specialists.

The following bibliography includes references cited in the above text as well as other sources that may be helpful in selecting educational materials and equipment.

Dale, E. and Chall, J.S. A Formula for Predicting Readability. Bureau of Educational Research, Ohio State University, 1948.

Eash, M.J. Talmage, H. and Walberg, H.J. Evaluation of Instructional Materials TM Report No. 45. ERIC Clearinghouse on Tests, Measurements and Evaluation. Princton, N.J. Educational Testing Service, October, 1975.

EPIE Career Education S*E*T, Volume 1, No. 65, 1975.

EPIE Educational Materials Report, No. 85m, 1978.

EPIE Educational Product Report, No. 65, 1974.

Fry, Edward. A Readability Formula that Saves Time. Journal of Reading. April, 1968, pp. 513-516.

Harris, A.J. and Sipay, E.R. How to Increase Reading Ability. New York: David McKay, 1976.

Selection Guidelines: School Library Resources, Textbooks, Instructional Material. Albany: New York State Education Department, Bureau of School Libraries.

"Standard Criteria for the Selection and Evaluation of Instructional Materials." National Center on Educational Media and Materials for the Handicapped. The Ohio State University, Columbus, Ohio.

Woodbury, Marda. Selecting Instructional Materials. Fastback 116. Bloomington, Indiana. Phi Delta Kappa Educational Foundation, 1978.

CHAPTER 6
EVALUATING INPUTS INTO IV-B SUB-PROJECTS -- MATERIALS AND EQUIPMENT

The evaluation of materials and equipment requires special considerations and techniques. Evaluations of materials designed for instructional purposes can be based on content analyses of the materials and on perceptions of users such as students, teachers and supervisors. In addition library materials can be evaluated by patterns of use. Equipment may be evaluated by means of ongoing judgements of performance and quality, records of use, and judgement of usefulness.

The section of this Handbook entitled Assessing Participant Attitudes Towards Educational Programs, pp. 46-52 describes techniques which can also be used to assess the attitudes of teachers, students and supervisors towards IV-B materials and equipment (e.g., satisfaction with the performance quality, interest level, attractiveness of these resources). The present chapter consists of a description of techniques for keeping records of circulation and frequency of use of educational materials and equipment. These techniques have been modified, based largely on suggestions made by respondents to the survey of IV-B evaluation methods, to permit inferences about the conduct and quality of educational programs.

Records of Circulation for Educational Materials

Circulation statistics are records of the patterns of use of educational materials. On the whole, if the procedures for collecting such information are systematic and properly implemented then the data obtained tend to be highly reliable. Personal judgements are not likely to cause serious anomalies in this type of information.

Validity, on the other hand, can be a serious concern. In assessing validity, the first question must be "What purpose are circulation statistics to serve?" If the purpose is simply to determine how often various materials were used, there would be little question that the method is appropriate to the purposes. It is not uncommon however, to find inferences like the following based on such records:

- "The high frequency of use of the Title IV-B equipment indicated that they were of acceptable quality."

- "Soon after the Remedial Reading sub-project began, students who were participating in the sub-project began to take more books out of the library and continued to do so for the remainder of the school year. This indicates that their interest in reading has increased as a result of the sub-project."

In both cases the inferences made on the basis of records of use could easily be invalid. Methods used to collect such information rarely have provisions for assesing the extent to which factors other than the observed ones are responsible for the results. In other words, these other possible factors have not been "controlled." In the first example there was no control to determine whether or not equipment was used so frequently because use was required or because no other equipment was available. In the second case, no controls were established to determine what proportion of the circulation increase observed was due to teacher assignments rather than to student interest. When the evaluation question to be answered is asked before an instrument is chosen or applied, then there will be an opportunity to design the instrument and control conditions so as to collect data that will permit valid inferences.

Persons responsible for IV-B sub-projects who responded to the survey noted that:

1. Circulations statistics do not reflect the use of materials within a media center. In certain programs within-center usage might constitute a substantial proportion of total use.

2. Patterns of use of materials and equipment often are not meaningful when considered for periods of time up to a year. Sometimes the examination of statistics for several years is necessary in order to make useful interpretations.

An Instrument for Collecting Evaluation Information Based on Circulation Records

The question of the appeal that educational materials and equipment have for students is an important one. The feelings of students concerning school library and instructional resources and the relationship between these feelings and academic success are both worth investigating.

Figure 7 displays an instrument for collecting information on the appeal and appropriateness of materials as well as circulation information. It is a synthesis of suggestions made primarily by some of the librarians who supervised Title IV-B subprojects. It is particularly appropriate to school library and classroom books that are available for circulation.

Date_____ Class_____

Book Title (This may be filled in by the teacher or librarian)_____

I Borrowed This Book Because:

_____It looked nice
_____It looked interesting
_____A friend told me it was a good book
_____My teacher told me it was a good book
_____My teacher asked me to borrow it
_____I don't remember why I borrowed it.

The Book Was:

_____Boring
_____O.K.
_____Interesting
_____Very interesting.

The Book Was:

_____Too difficult to read
_____Difficult, but worth reading
_____Easy to read
_____Too easy.

I Read:

_____All of the book
_____A large part of the book
_____Just a few pages
_____None of it.

Figure 7. An evaluation instrument for circulated materials

This method involves the use of special cards placed in reading materials which are part of a IV-B sub-project. A request made by the librarian or printed on the cards encourages the borrower to supply the desired information. The wording should be adjusted to fit the student's grade level. This type of information permits reasonable inferences about the appropriateness and interest value of reading materials and can help answer questions about the extent to which increased circulation was due to teacher assignment or the content and/or attractiveness of the materials.

Evaluating Equipment in Educational Settings

The primary expectations for articles of equipment (motion picture and slide projectors, cassette recorders, etc.) is that they be at acceptable levels of performance, usefulness, and durability within the context of user requirements.

The expected performance is relative to the nature of the equipment under consideration and necessarily involves different criteria for cameras than for audio-cassette recorders and computer terminals. Methods and criteria for conducting formal evaluations on specific types of equipment are provided in a series of publications produced by EPIE (Educational Products Information Exchange) Institute.

The usefulness of a particular piece of equipment can be assessed to some degree by the frequency of its use and the demand for its use. Complexity of instrumentation can often be a deterrent to maximal use, but this is a problem that can usually be solved by inservice training. Compatability is another important criterion for evaluation of usefulness. A piece of equipment which is not compatable with the existing set of media resources would tend to be used less often. Concerning durability, a study conducted by EPIE on equipment malfunction (EPIEgram: Equipment, Vol 7e, No. 2, October 23, 1978) indicates that while almost 18 percent of malfunctions can be attributed to poor equipment design and 31 percent to normal wear and tear, a disturbing 33 percent are related to mistakes made by users of the equipment. This is another important reason for providing training for users and for collecting information as to the need for such training.

A useful method for keeping track of these aspects of equipment quality and use is to accompany borrowed equipment with information forms, such as the one displayed in Figure 8. Such forms are filled out by the user and returned along with the equipment.

The key to the successful use of such forms is that they be as easy to complete as possible. A multiple choice format is best, leaving space for user comments.

The Evaluation Report on Equipment Usage, Figure 9, shows a possible analysis of one of these survey forms when the responses to item three of the Survey of Equipment Use Form are coded with values from 1 to 5 going from poor to excellent and the responses to item four are coded 1 to 4.

The Average Performance Rating is the mean of responses to item three.

The Complexity Rating is the ratio or reports operating complexity (question 2) to the total number of times the unit was used. If the ratio is relatively high one may suspect that complexity is interferring with potential use.

The Overall Reaction is the mean of ratings given to item four.

Survey of Equipment Use

Date _____

Type of Equipment:

Serial Number:

(This information may be filled out by media personnel to save time for the user.)

Borrower:

Used For:

Please fill out the following form before or when returning borrowed equipment to the Media Center.

1. Was any part of the equipment broken or malfunctioning?

   ☐ Yes

   ☐ No

If yes, please describe _____

_____

2. Did the operational complexity of this piece of equipment make it difficult to use if efficiently?

   ☐ Yes

   ☐ No

3. What was the level of performance of this piece of equipment?

   ☐ Exceedingly poor ☐ Fair ☐ Excellent

   ☐ Unsatisfactory ☐ Good

Comment _____

_____

4. Would you use this piece of equipment again?

   ☐ No

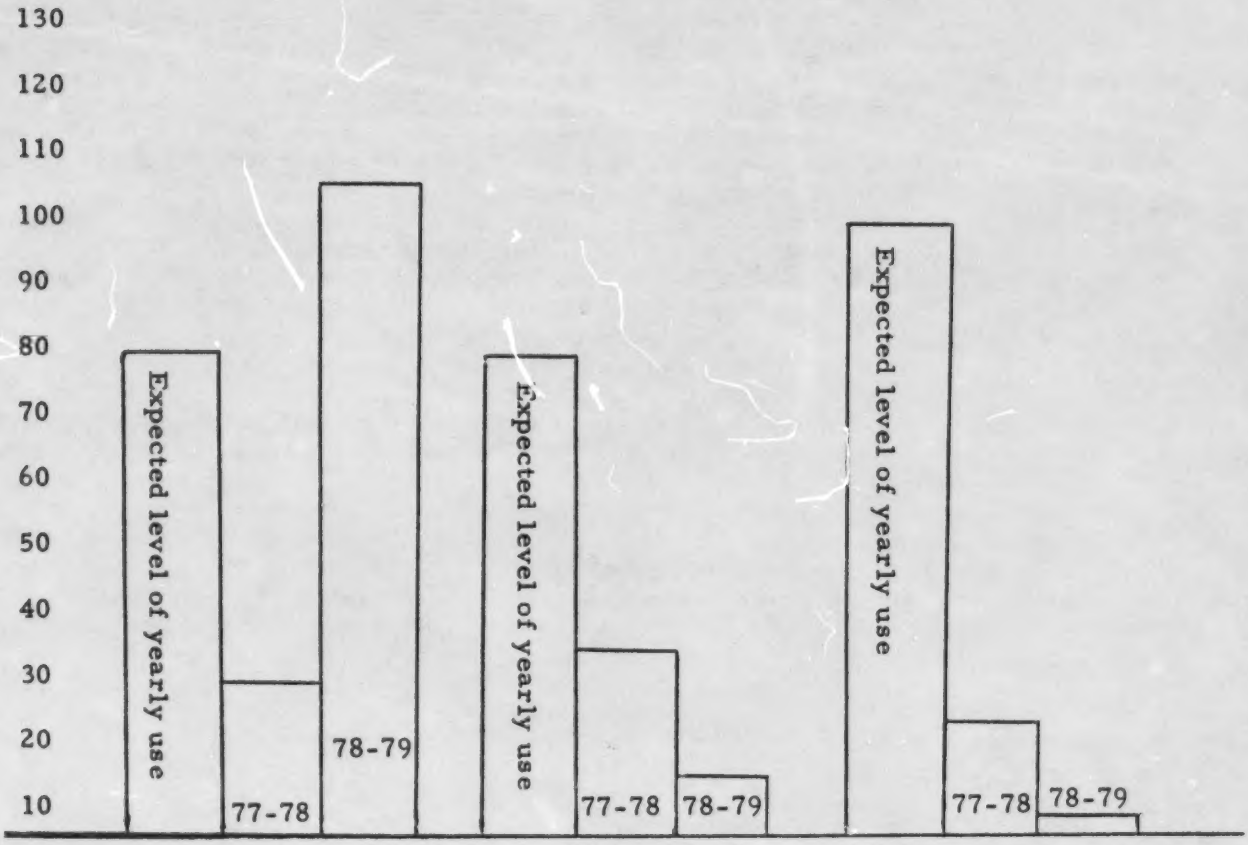
   ☐ Only if necessary

   ☐ Yes

   ☐ Eagerly

Figure 8. An instrument for evaluating equipment in educational settings.

Evaluation Report on Equipment Usage
Part 7 - Video Cassette Units



Equipment Code	C230		C234		C411	
School Year	1977-78	1978-79	1977-78	1978-79	1977-78	1978-79
Number of Malfunctions	10	2	23	15	2	2
Average Performance Rating	4.1	4.5	4.2	4.3	2.1	2.3
Complexity Rating	.80	.19	.75	.20	.30	.30
Overall Reaction	2.9	3.5	1.8	1.1	1.2	1.3

Figure 9. Analysis of Data from "Survey of Equipment Use"

Results tabulated and analyzed in such a fashion can provide information which goes beyond mere frequency of use and can in fact help to explain a pattern of use. For instance, the 1977 evaluation of C230 indicated an overly high record of malfunctions and a high complexity rating along with frequency of use below expected levels. This suggested that need for inservice training on the possibilities and most efficient use of this unit. The 1978-79 ratings show dramatic improvements following training. The ratings and usage information in C234, a similar piece of equipment, which was also a subject of inservice training might have been expected to show a similar improvement, but in general they did not. The complexity rating was reduced so the inservice workship was evidently successful; the problem seems to be that the unit either sustained a serious injury at some early point or is defective in some way, because the number of malfunction reports is disproportionately high. For this reason it did not attain the confidence of past and potential users. C411 on the other hand appears to be disfavored because of its low performance quality.

Many school districts maintain record on frequency and nature of equipment malfunction, and a running count of repair costs on each piece of equipment as well as a record of frequency of use.

In summary, it is important to remember that while ongoing educational program records such as circulation and frequency-of-use statistics tend to be objective and reliable, they usually cannot serve as a basis for evaluation inferences unless the instrument used to collect information is designed to answer specific evaluation questions.

CHAPTER 7
EVALUATING THE IV-B PROCESS -- ACHIEVEMENT OF INSTRUMENTAL OBJECTIVES

Instrumental objectives are expectations for changes in program conditions. These can involve either the improvement of the quality of an existing program or the introduction of a new activity or product into that program.

The following are examples of questions that might serve as the basis for the assessment of progress towards the achievment of instrumental objectives:

Was X service or facility made available to more students than in the past?

How often did students participate in X?

What was the typical or accumulated duration of their participation?

To what extent were students made aware of the sub-project's resources, goals and rationale?

To what extent were faculty prepared to implement the sub-project?

To what extent was the sub-project implemented according to schedule?

To what extent was supervisory and administrative support available?

These and a number of other similar questions could serve as the basis of an evaluation related to instrumental objectives. There are three main methods for collecting information to answer such evaluation questions:

1. Attitudes of participants (i.e., students and staff) towards some aspect of the educational program
2. Observational procedures
3. Ongoing program records.

This chapter contains sections on the assesment of participant attitudes and observation procedures. The use of ongoing program records in educational evaluation is discussed in Chapter 6, Evaluating Inputs Into IV-B Sub-Projects -- Materials and Equipment.

Assessing Participant Attitudes Towards Educational Programs

This section is a discussion of techniques for measuring attitudes towards some aspects of an educational program. The term "attitudes" refers to systematic tendencies of individuals or groups to think, feel, perceive, and believe in a certain way towards some referent concept or object. Underlying the concern for attitudes towards a program is the belief that the participants attitudes (e.g., students and teachers) will affect their performance and are therefore important to consider.

The methods presented here involve the use of rating scales to record the attitudes of participants towards program activities. These methods are selected from a wide range of techniques that can be used to assess attitudes. They are chosen because they can be implemented and interpreted in a relatively straightforward manner and do not require complex analysis techniques when used in their simplest form. The references listed at the end of this section provide a wide range of methods, as well as many techniques for design and analysis associated with each method.

Two types of rating scales for assessing attitudes will be discussed. One is called a numerical rating scale or Likert scale; the other is called the semantic differential. Both yield scores which can be interpreted in terms of the direction and magnitude of persons' attitudes towards some aspect of an educational program.

Numerical Rating Scales

Numerical rating scales consist of sets of items each composed of a statement and a graded dimension with provisions for giving an attitudinal response to the statement on that dimension, as in the following example:

The photography lessons were clearly presented.

SA	A	U	D	SD
Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree

The dimension (of agreement in this case) may be:

1) Presented under or to the right of, each statement with provisions for circling, checking or copying ratings, or
2) listed at the top of the page and a space provided next to each statement for entering ratings.

Although the agreement-disagreement continuum is the most common, others are possible. For example:

Supervisors have been available to support implementation of the sub-project.

Never	Rarely	Sometimes	Frequently	Always

Each rating is given a numerical value, typically 1 to 5 for a 5-point scale or perhaps -2 to +2 for those who feel comfortable working with negative numbers. One end of the scale (usually the highest) is chosen to represent the positive or desirable end of the dimension. Sometimes for convenience the numbers themselves are placed on the rating form.

The simplest way to analyze the data from such scales is to find the average rating across individuals for each statement. (1) When the rating format is identical for each statement (e.g., each dimension goes from SA to SD), (2) There is a common meaning to the statements, and (3) The attitudinal referent is common to all statements (e.g., a particular class, set of curriculum materials, subprojects, etc.) then the ratings may be summed and averaged across statements, as well as across individuals for each statement.

For example, let us consider an attitude scale concerned with assessing faculty perceptions of the extent and quality of supervisory support for the implementation of a Title IV-B sub-project. The instrumental objectives related to this support are translated into a set of statements, which become the Title IV-B "subscale" of a faculty questionnaire, as shown below:

a. Teachers were given an important role in decision making process

SA	A	U	D	SD

b. Teachers were made aware of the sub-project rationale and goals

SA	A	U	D	SD

c. Teachers were made aware of available resources

SA A U D SD

d. Teachers were adequately prepared to make use of sub-project materials and equipment

SA S U D SD

e. Supervisory and administrative support was available when needed

SA A U D SD

f. Supervisory and administrative support was helpful

SA A U D SD

Sets of items dealing with a common referent, as in this example, are referred to as subscales. Since it is usually difficult and even questionable to judge some person, resource, or activity from ratings on a single statement, it is desirable to develop subscales for each object of evaluation. The presence of sets of carefully crafted items, rather than lone items for each program facet, will help to counterbalance errors due to weaknesses of individual items and thereby improve both the validity and reliability of an instrument. On the other hand, average scores for the individual items can often provide valuable suggestions concerning program facets which may need attention.

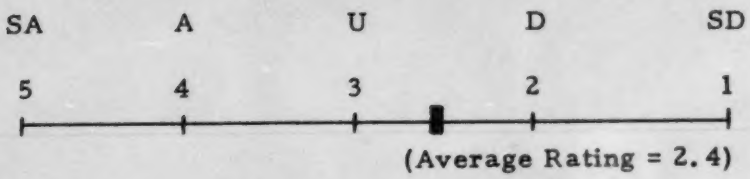
There are several ways in which the tabulated scores from attitude rating scales may be displayed.

Extending the example of the supervisory support evaluation just presented to the analysis stage, an initial step would be the tabulation of frequency counts and percent of total responses for each response category followed by a calculation of the average ratings obtained for each item as shown below:

Teachers were given an important role in decision-making processes.

SA	A	U	D	SD	Total
7 (4%)	4 (7%)	14 (25%)	24 (43%)	10 (18%)	56 (100%)

OR



The average ratings for individual items take on added meaning when viewed in the context of responses to the subscale as a whole as shown in figure 10.

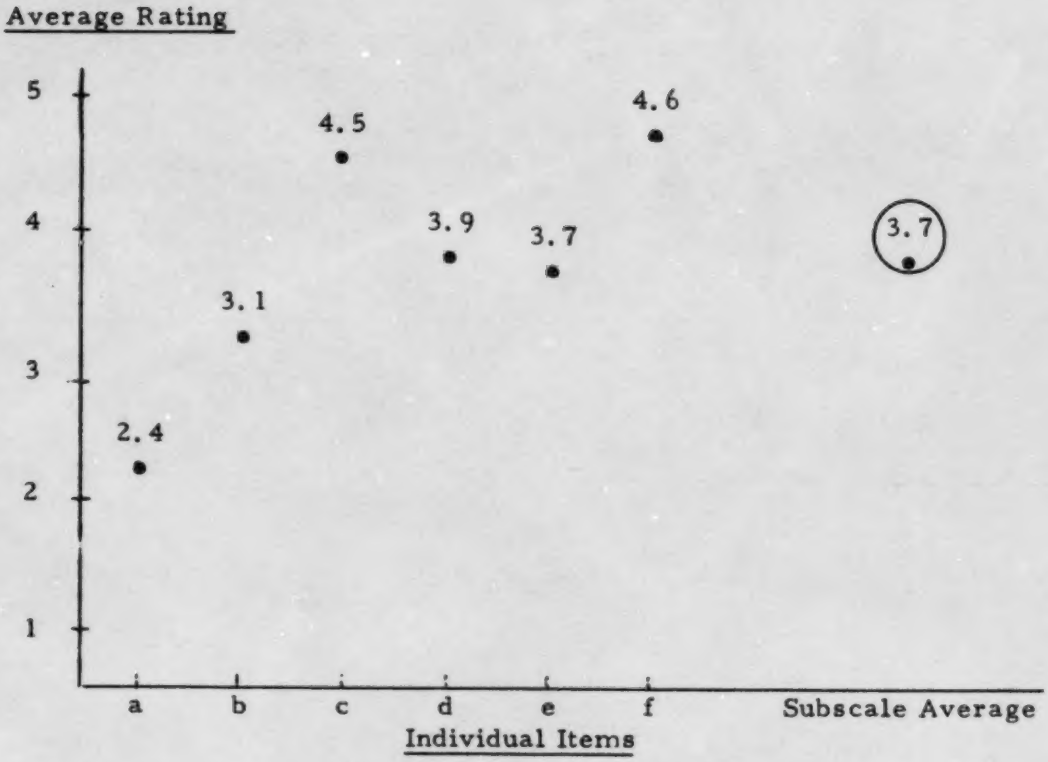


Figure 10
Items on IV-B Subscale From Faculty Attitude Questionnaire

These results suggest that while teachers tended to feel that they may not have been included sufficiently in the conceptualization aspects of the sub-project, on the whole they perceived the supervisory conduct of program implementation to have been satisfactory.

Prior specifications of expected levels for the ratings, or of levels which could be considered minimally satisfactory will greatly aid interpretation of such analyses and displays.

The Semantic Differential

The semantic differential is a method of measuring attitudes based on the notion that attitudes towards various objects of concern can be expressed in terms of polarized dimensions (e.g., good-bad, strong-weak, bright-dark, etc.) and that, while there may be differences in the connotations that individuals have for objects, that there are also common elements of meaning that can be expressed in terms of locations on these dimensions.

Imagine that a high school language arts program, including instructional materials, services, and hardware, has been purchased and installed in a district high school.

Four aspects of this program are seen as crucial to its success, and a decision is made to assess faculty and student reactions towards these facets. The semantic differential technique might be used to serve this purpose in the following way:

1. Identify the concepts to be rated. In this case they are the four program facets mentioned above:

   -- Laboratory units
   -- Curriculum materials
   -- Tests
   -- Support services (provided by the program producer).

2. Choose adjective pairs which will make up the extreme ends of bipolar scales.* The choice should be on the basis of relevance to the issues at hand. Bipolar scales for the present purpose might include:

Good-Bad
Clear-Confusing
Valuable-Worthless
Relaxed-Tense
Pleasant-Unpleasant
Organized-Disorganized
Useful-Useless
Realistic-Impractical
Appropriate-Inappropriate
Easy-Difficult
Limited-Comprehensive
Encouraging-Frustrating
Warm-Cold
Dull-Lively
Friendly-Unfriendly
Boring-Interesting
ETC.

*(The development and validation of bipolar scales have been the subject of a great deal of research by the originators of the semantic differential technique. Instrument developers should be sensitive to the fact that what may appear to them to be a bipolar dimension may not in fact be such for some or all of the persons who will be filling out evaluation forms. One solution to this problem is to make use of dimensions such as those used by Osgood, Suci and Tannenbaum (1975), which have already been validated).

3. Prepare a response sheet according to the following format, with a concept at the head and the complete set of bipolar adjectives under each concept:

Support Services

Warm	:___:	:___:	:___:	:___:	:___:	:___:	:___:	Cold
Boring	:___:	:___:	:___:	:___:	:___:	:___:	:___:	Interesting
Clear	:___:	:___:	:___:	:___:	:___:	:___:	:___:	Confusing

The response sheet should include instructions to respondents to place a mark at the place on each dimension whose location between the two extremes expresses their feelings towards the concept. A seven-point scale is typical. For younger children it is usually found appropriate to use a five-point scale and to limit the number of dimensions under each concept.

4. Assign a point value to each division on the scales. In this example, the most straightforward assignment would be values of 1 to 7 with high values most commonly standing at the "positive" end of the scale. A simple analysis would involve computing group means for each adjective pair and plotting them as in Figure 11.

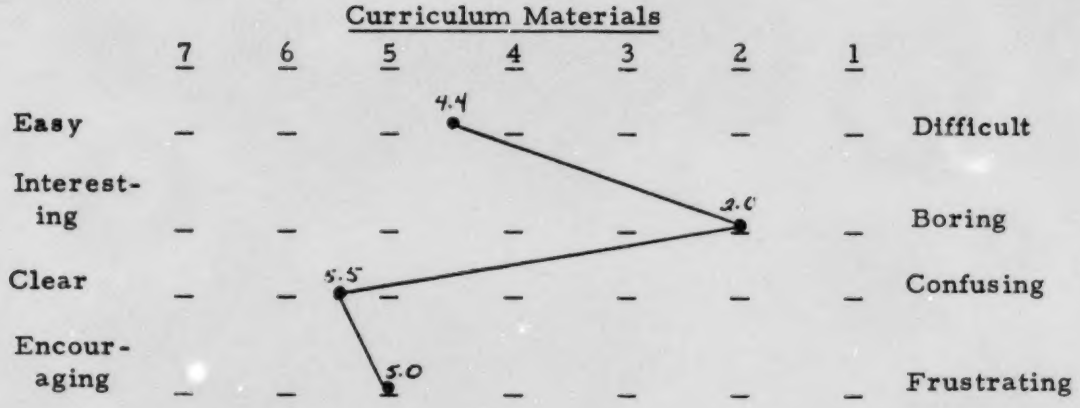


Figure 11. Group Means for a set of Semantic Dimensions

One can also tabulate the frequencies given each rating, as in Figure 12, and plot these on a frequency polygon, as shown in Figure 13, for the tabulated results drawn from ratings given to curriculum materials.

Tabulation

	Value Assigned						
	7	6	5	4	3	2	1
Easy	11	𝍸 ///	𝍸 11	11	𝍸 11	𝍸	___ Difficult

Average Rating
4.4

Figure 12. Tabulation of Responses to a Semantic Dimension

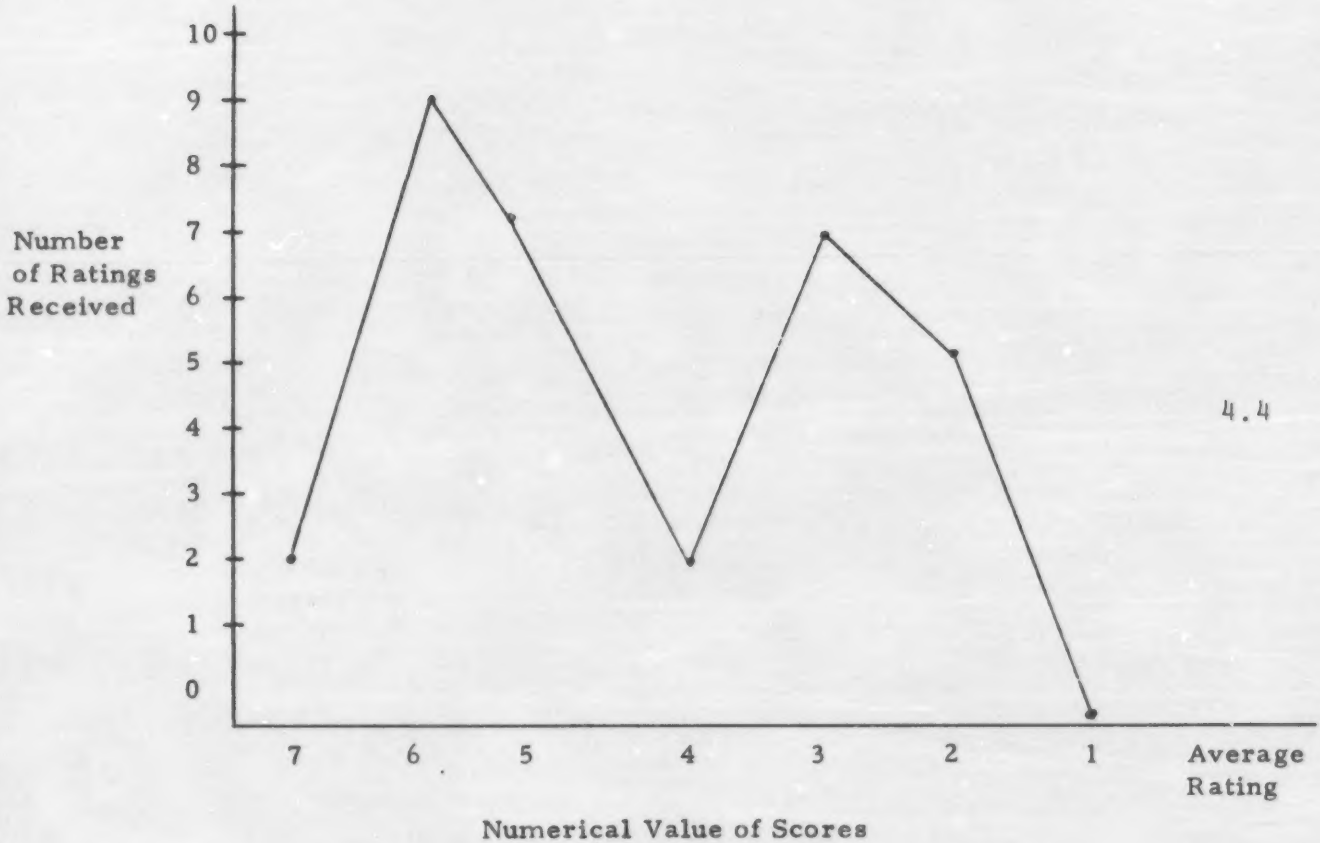


Figure 13. Frequency Polygon Display of Ratings on the Easy-Difficult dimension

It is recommended that both types of analyses be conducted, as they provide different kinds of information. An average rating of 4.4 on the difficulty dimension, for example, does not reflect the fact that there appear to be two distinct groups of respondents on this dimension; one group finds the materials relatively easy and the other finds them relatively difficult. Further investigations might reveal the characteristics of pupils in each of these groups.

Validity and Reliability of Attitude Measures

Providing Instructions That Will Accompany Instruments

As with all pencil and paper tests, it is important to accompany rating scales with clear instructions describing their purpose and rationale and explaining how to fill out the instrument. Instructions should be supplied with examples of how to complete the instrument properly. The instructions and format of an instrument are often as much in need of field testing as its content, and valuable information on their effectiveness can be obtained by observing and questioning a sample of respondents as they try the instrument out.

"Response Sets"

Many individuals tend to respond to attitude scales in ways which bear no direct relationship to the attitude being studied. For example, some persons tend to mark extreme ends of a rating scale, while others tend to concentrate their ratings on the center of the continuum. This tendency is called a "response set" and it can have a harmful effect on the validity of the information derived from rating scales. Instructions requesting respondents to use as much of the full range of a scale as possible can sometimes counteract such tendencies.

"Halo Effects"

This term refers to the tendency on the part of respondents to rate some referent (i.e., some aspect of the educational program) in a constant direction based on a global impression of that referent. For example, a student's perception of school activity which is too difficult for him or her, might result in a tendency by the student to rate that activity as "boring," "useless," "inappropriate," and perhaps even "illogical" as well. In cases where a general positive or negative impression towards some referent is desired, this tendency may actually result in desired effects, but for the most part it will tend to obscure distinctions which are important to the evaluator, and thereby reduce the validity of obtained information.

Respondent Group Size

Inferences based on affective measures tend to be less reliable and valid for individuals than they are large groups. This is so partially because the range of responses derived from groups tends to balance out some of the effects of individual response tendencies which may seriously bias the scores of those individuals. An isolated group of responses, derived from scales, such as the ones described based on one individual, or a few individuals, will typically be uninterpretable, unless it can be related to the patterns of an appropriate norm group.

The following sources can provide further information on the development of instruments for assessing attitudes, as well as on techniques for analyzing and interpreting obtained results:

Anastasi, A. Psychological Testing. (4th ed.) New York: Macmillan Publishing Co.,1976.

Edwards, A.L. and Porter, B.C. The Affective Domain. Washington, D.C.: National Special Media Institutes, Gryphon House, 1972.

Guilford, J.P. Psychometric Methods, (2nd ed.) New York: McGraw Hill, 1965.

Henerson, M.E., Morris, L.L. and Fitz-Gibbon, C.T. How To Measure Attitudes. Beverly Hills: Sage Publications, 1978.

Osgood, C., Suci, G., and Tannenbaum, P. The Measurement of Meaning. Urbana, Illinois: University of Illinois Press, 1957.

Payne, D.A. The Assessment of Learning; Cognitive and Affective. Lexington, Massachusetts: Heath, 1974.

Shaw, M. & Wright, J. Scales for the Measurement of Attitudes. New York: McGraw Hill, 1967.

Observational Procedures

Observational procedures require that one or more individuals direct their attention in a systematic fashion to some activity within an educational program. In order for the information obtained to be useful for later decision making, the observational procedures must be structured to achieve their purpose and yet be easy to use. Appropriate instruments for this purpose include tally sheets or checklists and behavior coding systems. These often require that some behavior of teachers, students, or both be categorized. A good way to derive these categories is to list features of student and/or teacher actions which are either desirable or undesirable with respect to the achievement of program goals.

For example, it might be determined that the greatest benefit from a particular structured learning program is obtained when the teacher:

(a.) has students take responsibility for implementing specified portions of learning program activities
(b.) devotes seventy percent or more of class time to providing individual assistance to students
(c.) maintains current records on student progress
(d.) assures that each student is following the correct sequence of progress through the learning program.

Based on these categories, checklists and observation schedules can be developed to keep ongoing records of the extent to which the specified expectations are being met.

On the instrument itself the categories may be represented by a column to be checked, or by a code number or letter to be recorded in a prescribed manner for each instance of that behavior (see figure 18). While it is important to categorize and code relevant features beforehand, room should be left for comments by raters in case relevant program features which were inadvertantly left out of the classification scheme are discovered. This should lead to an improved instrument for later rounds of observations.

For the instrument to be effective, it is essential that these categories be mutually exclusive and that the behaviors of interest be clearly assignable to categories. Such clarity is often difficult to assure before an actual tryout. A practice run in which at least two persons simultaneously fill out the instrument while observing the same setting will give an indication of how clear and simple the system is. More formal means for assessing the reliability of such instruments by testing the degree of agreement between different observers can be found in the Guilford and Siegel texts cited in the bibliography. Training and rehearsal are often necessary to bring interobserver reliability up to an acceptable level. In general, the fewer the inferences the observer has had to make in filling out an observation instrument, the greater the interobserver reliability will be.

The choice of criterion behaviors for observation purposes can result in a problem which often requires a trade-off between concerns for validity and reliability For instance, if an observer is asked to count the number of instances of "constructive participation in classroom presentations," differences of opinion of the part of observers as to what constitutes constructive participation may result in discrepant interpretations of the same classroom occurrences. This would result in reduced reliability of observation scores. The typical solution to the problem is to make sure that the

criterion behaviors are as clearly defined and specific as possible. For "constructive participation," we might count the number of instances in which a student raised his/her hand instead of calling out in class. There would be little question as to when this behavior is or is not taking place, but then another problem surfaces. There may be important instances of constructive particpation in which a student has not raised his/her hand, and in fact there may be situations and settings where hand raising can be dispensed with without destroying the productivity of the educational interchange. Hand-raising itself, then, is not a universally appropriate criterion of constructive participation and the validity of an instrument would be reduced if it depended heavily on such a criterion.

It is important to recruit observers who are disinterested as possible in the outcomes of a program. When observers have a stake in the outcomes of the activity they are studying the information obtained is potentially open to charges of bias.

As well as recording categories of behaviors, observers may also be asked to make judgements on rating scales similar to those discussed in Assessing Participant Attitudes Towards Educational Programs.

Observation methods have the advantage of reducing the data-collection burden on teachers; on the other hand, the presence of observers in their classrooms can make some teachers uncomfortable, particularly the first few times it occurs.

Another essential step in planning observational procedures is choosing an appropriate sample of situations to observe. Information needs which the evaluation must satisfy will help decide:

- Which situations and behaviors will be observed
- How often observations will be made
- How long they will last
- Who will be observed and/or where the observation will take place.

The following example, based on the individualized learning program situation described above, suggests how a scheme for collecting, analyzing and displaying data to evaluate such a program might be designed.

First, consideration is given to the type of information that will be valuable for indicating the extent to which the four expectations are being realized:

> For objective (a) it is found that six aspects of program implementation may profitably be turned over to students. It is decided that a simple tabulation of these events at different points during program implementation will provide the desired information.
>
> For objective (b) it is decided that observers will estimate the proportion of time that teachers devote to providing individualized assistance to students and that several interobserver comparisons will be made to see if these estimates are reliable.
>
> For objective (c) a procedure is set up for allowing classroom observers to assess through an examination of records the number of days or class periods that the teacher is behind in recording individual records of student progress onto a master planning book.
>
> For objective (d) a method is devised for randomly selecting five students during each observation and assessing how many are at the correct point in the sequence based on their individual learning program records.

Class _____

Extent to which Students Assumed Responsibility for Specified Learning Program Activities

Student Responsibilities	Week of-- 2/4	2/11	2/18	2/25	3/3	3/10	3/17	3/24	3/31	4/7	4/14	4/21	4/28	Total Point Values
1. Distribution of Materials	◪	◪	■	■	■	■	■	■	◪	SPRING RECESS	◪	■	■	20
2. Maintaining Materials in Correct Order			◪	◪	◪	■	■	■	■		◪	■	■	16
3. Recording Amount of Time Spent on Various Activities			◪		◪		◪	◪				◪	◪	6
4. Recording Scores	◪	■	■	■	■	■	■	■	■		■	■		21
5. Preparing Reviews of Materials for Other Students	◪	◪		◪	◪	◪	◪	◪	◪				◪	9
6. Providing Assistance to "Team Members".			◪		◪	■	■	■	◪			◪	◪	10
TOTAL Point Values	3	4	7	6	8	8	10	10	7		4	8	7	

KEY

□ - Students have not taken responsibility (0 points)

◪ - Some indication that students are taking responsibility (1 point)

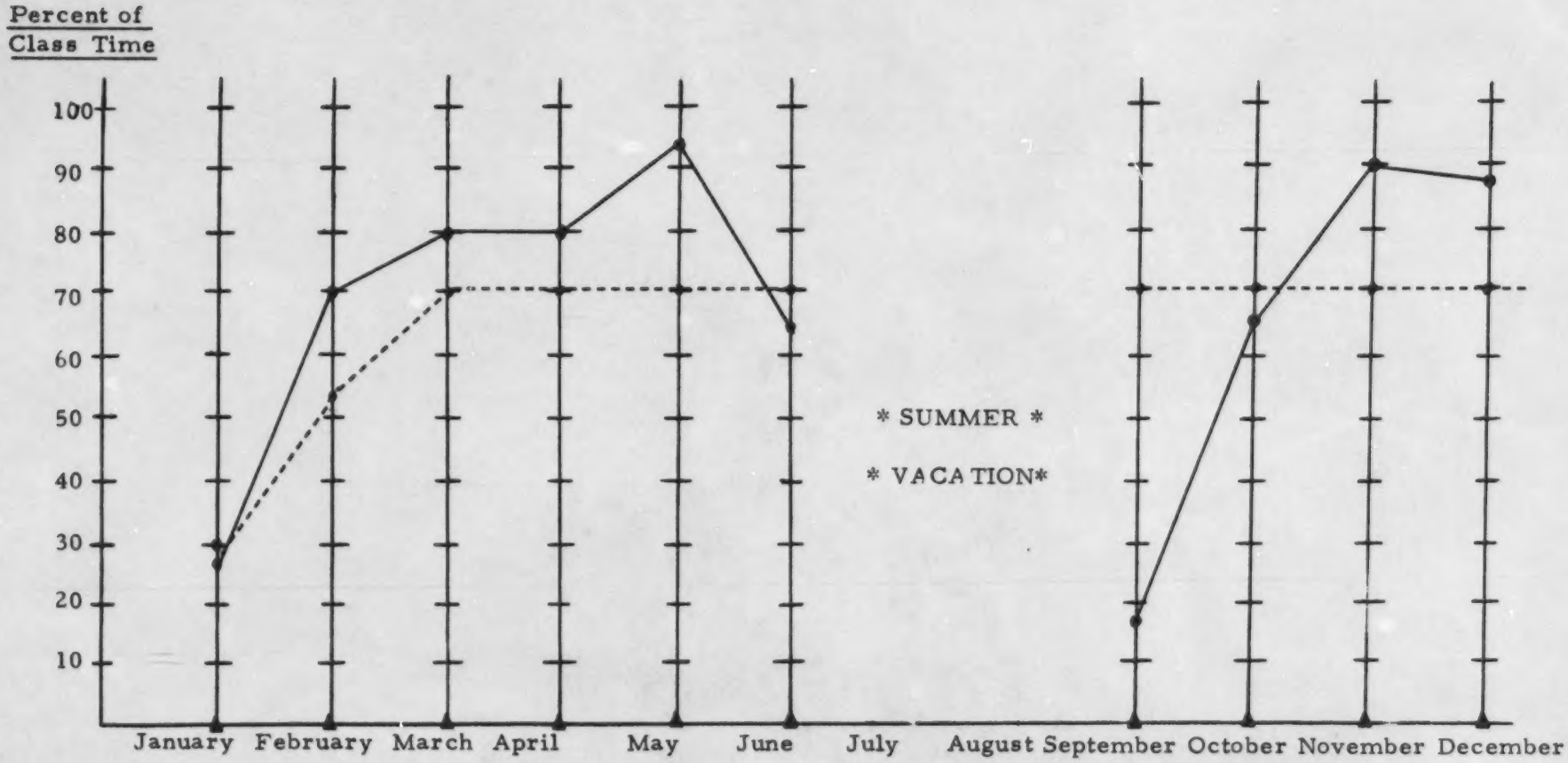
■ - Students are successfully implementing prescribed duties (2 points)

expectation

By 2/18 classes will be scoring 1 point on each activity and by 3/10 will reach and maintain a 2 point level

Figure 14. Analysis Sheet for Objective "a."
(Students will take responsibility for implementing portions of learning program activities)

Proportion of Class Time Devoted to Individual Assistance



---------- Percent of time teacher is expected to spend providing individual assistance to students.

_____ Percent of time actually spent by teacher providing individual assistance to students.

Figure 15. Objective "b."
Teacher devotes seventy percent or more of class time to provide individual assistance to students

Class_____

Number of Class Periods Needing to be Recorded
in Order to Bring Master Records up to Date

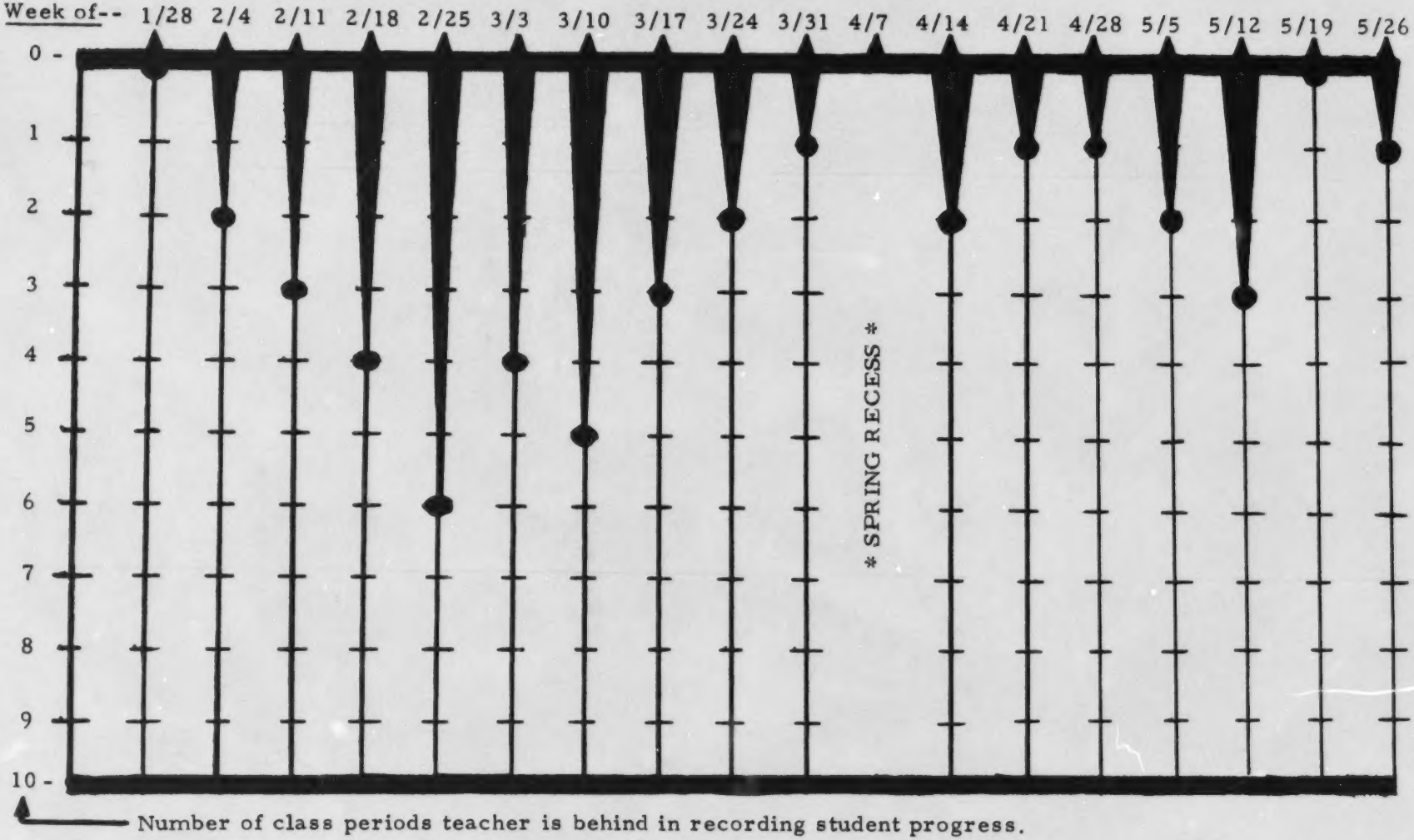
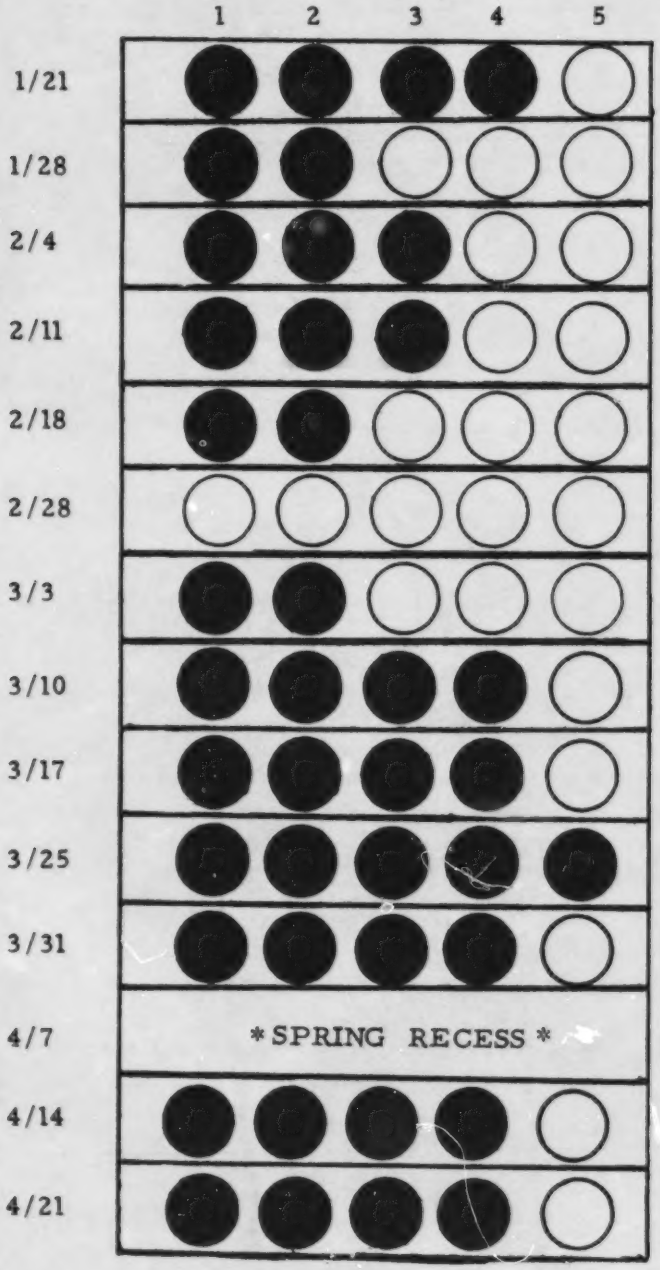


Figure 16. Objective "c"
Teacher maintains current records on student progress.

Number of Students Found to be Working at the Point in the Learning Program Which is Appropriate for them
(out of a different sample of five examined each week)



KEY

○ = one student

● = student found to be working at the appropriate stage in the sequence

Figure 17. Objective "d"
Teacher assures that each student if following the correct sequence of progress through the learning program

86

CLASSROOM REPORT FORM - TITLE IV-B LEARNING PROGRAM

Class Elsie Doe

Date 3/10

Observer G. Samsa

a.

		☐	◪	■
1.	Distribution of materials	—	—	✓
2.	Maintaining materials in correct order	—	—	✓
3.	Recording amount of time spent on various activities	✓	—	—
4.	Recording scores	—	—	✓
5.	Preparing reviews for other students	—	✓	—
6.	Providing assistance to team members	—	✓	—

b.

Percent of class time teacher spent providing individual assistants to students

10 20 30 40 50 60 70 (80) 90 100

c.

Number of class periods needing to be recorded in order to bring master records up to date: 3

d.

Number of students found to be working and the appropriate stage in the sequence (out of 5 student records checked): 3

Figure 18.

Instrument for collecting data on program implementation

It is decided that weekly visits, each involving one 45 minute class period of observations, will be made followed by an examination of teacher and student records which will take about 20 minutes. Expected levels of performance are set for each objective and a plan for examining the final results is devised. Information on objectives (a), (b), (c), and (d) could be displayed as in the charts shown in Figures 14, 15, 16, and 17 respectively, with criterion levels and projected results indicated for each objective.

An instrument such as the one shown in Figure 18 could then be designed to collect this information.

Variations of this evaluation procedure can also be used to observe aspects of program implementation such as instructional styles and interactions between students and teachers. For instance, an observation instrument might be constructed based on the sequence of procedures an instructor is expected to follow in the use of a learning kit, or the proportion of time devoted to particular activities which make up an educational program.

The following sources provide discussions of observational procedures and samples of numerous instruments that have been used for these purposes. The Guilford and Siegel texts are sources of techniques for assessing interobserver reliability.

Borich, G.D. and Made, S.K. Evaluating Classroom Instruction: A Sourcebook of Instruments. Menlo Park: Addison-Wesley Publishing, 1977.

Guilford, J. Psychometric Methods. (2nd ed.) New York: McGraw Hill, 1965.

Heynes, R. and Lippit, R. Systematic Observational Techniques in G. Lindsey, (Ed.), Handbook of Social Psychology. Cambridge, Massachusetts. Addison-Wesley, 1954, Vol., I, p. 399.

Morris, L.L. and Fitz-Gibbon, C.T. How to Measure Program Implementation. Beverly Hills, Sage Publications, 1978

Siegal, S. Nonparametric Statistics for the Behavioral Sciences. (2nd ed.) New York: McGraw Hill, 1974.

Simon, A., and Boyer, E.G. Mirrors for Behavior: An Anthology of Classroom Observation Instruments. Philadelphia: Research for Better Schools, Center for the Study of Teaching, 1974.

Webb, E.J. Campbell, D.T., and Schwartz, R.D. and Sechrest, L. Unobtrusive Measures: Nonreactive Research in the Social Sciences. Chicago: Rand McNally,
1966.

CHAPTER 8
EVALUATING IV-B OUTCOMES -- ACHIEVEMENT OF PERFORMANCE OBJECTIVES

Performance objectives are expectations for changes in the attitudes, abilities and behavior of students. Three general methods relevant to outcome evaluation will be discussed in this chapter:

- o attitude measures as means of assessing affective growth
- o locally devised and standardized measures of achievement as means of assessing cognitive growth
- o observational procedures as means of assessing changes in student behavior.

Evaluating Affective Growth

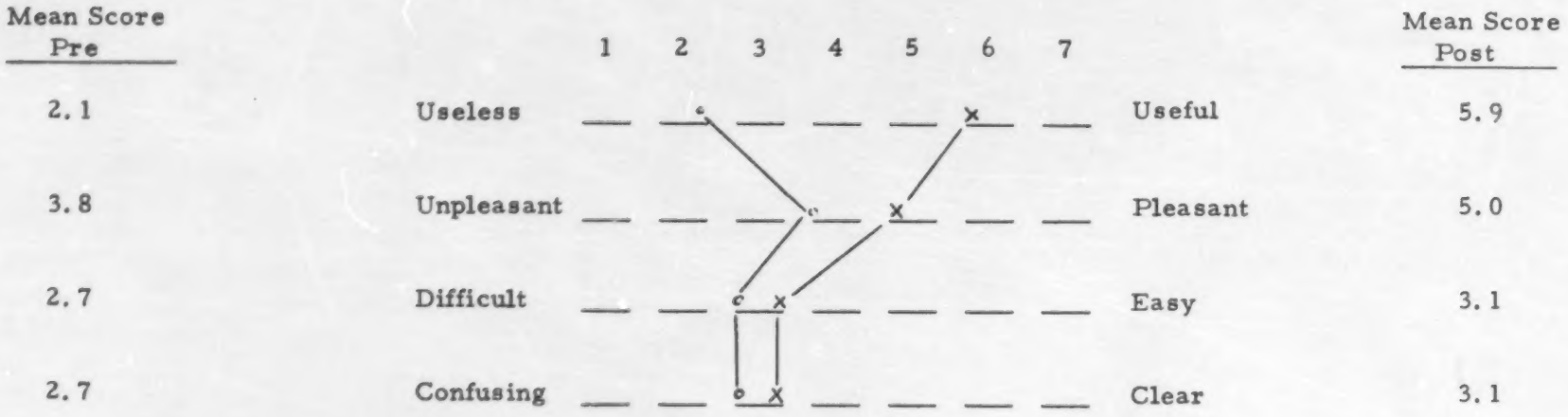
Evaluation of affective growth (the achievement of affective objectives) essentially involves the measurement of changes in attitudes. For this reason methods appropriate to this purpose are closely related to the methods of assessing attitudes discussed under the heading Assessing Participant Attitudes Towards Educational Programs p. 46-52. Since that section provides an introduction to basic issues and techniques for measuring attitudes, it also serves as a preface to the present section, which deals with the special considerations related to attempts to assess the achievement of affective objectives.

The difference between the two sections is primarily one of purpose rather than of method. The first section addressed how the attitudes of program participants toward some facet of an educational program could be used to evaluate the program. In this section the attitudes themselves are the subject of interest and the basis of performance objectives. For instance, the focus is on the development of children's attitudes towards reading (interest, curiosity, appreciation, commitment, etc.) in the course of participation in a given sub-project, rather than on their attitudes towards some facet of the sub-project, say, the quality of instruction, materials or equipment.

The following suggestions center on the use of attitudinal measures as methods for assessing the achievement of affective objectives. Expectations will typically be for affective growth or change of attitudes in a valued direction. For this reason, it is important to take baseline measures, both to be able to assess pre-post differences and to see if attitudes need changing in the first place. The final analysis should center on the statistics of growth. For example, when using the Semantic Differential Technique discussed above under Reports from Participants pre and post scores could be displayed together as in Figure 19.

Attitudinal measures differ from achievement measures in that they are not intended to represent the optimal performance of which individuals are capable under a given set of circumstances; instead they represent individuals' reports of their typical attitudes. Respondents can choose to modify their ratings in a way that they perceive will put them in a favorable light or will satisfy expectations of the person who is examining the results. For instance this might easily occur if students feel their responses are going to influence their course grades or their instructor's perception of them. This threat to validity can often be counteracted by allowing respondents to take attitudinal measures anonymously.

I Think History Is:


Mean Score Pre
1 2 3 4 5 6 7
Mean Score Post
2.1 Useless Useful 5.9
3.8 Unpleasant Pleasant 5.0
2.7 Difficult Easy 3.1
2.7 Confusing Clear 3.1

Summary of 7th Grade Attitudes Towards Social Studies
Survey Before and After the Title IV-B History in Action Subproject

mean pre-test scores

mean post-test scores

Figure 15. Method of Displaying Semantic Differential Ratings for Assessing Affective Growth

Evaluating Cognitive Achievement

Perhaps the most widely used instruments of evaluation in the field of education are achievement tests. Achievement tests are those which measure knowledge and skills, i.e., performance objectives in the cognitive domain.

The two most global decisions in the selection of achievement tests for program evaluation purposes are (1) whether to use locally designed or commercially produced tests and (2) whether the tests should be norm-referenced or criterion-referenced in nature.

Norm-referenced tests (NRT) are those in which the score of an individual takes on meaning in reference to the scores of other individuals who took the test. The performance of other individuals on the test thus serves as a norm or "yardstick" of achievement. The norms may be national, local, or even strictly classroom-based. Scores stated in terms of percentiles, standard scores, and grade-equivalent scores are norm-referenced. Most commercially produced standardized tests of achievement and ability (reading, basic skills, I.Q., etc.) are norm-referenced. Assigning grades "on a curve" is a technique of norm-referenced measurement.

Many educators are more interested in obtaining information on students' levels of attainment in some area of knowledge or skills than in comparing their performance to that of other test takers. Criterion-referenced tests (CRT) are designed to respond to this need. CRT's have a content which is directly related to asessing the achievment of specific objectives and by the existence of a "criterion level" or "cut-off" score which indicates mastery or acceptable performance. When teachers design or choose test items to assess the achievement of objectives (e.g., the ability to provide three alternative explanations for economic cycles) or set cut-off scores for acceptable performance (e.g., 90 percent of items correct on a spelling test) they are using practices which are characteristic of criterion-referenced measurement. The development of systematic means for determining test content and setting cut-off scores in CRT's is the subject of intensive investigation by educational researchers at the present time. The tests which accompany commercially produced learning kits and programs are typically CRT's.

Standardized norm-referenced tests of achievement and ability (e.g., Metropolitan Achievement Tests, Stanford Achievement Tests, etc.) are usually too global in content and scope for assessing the objectives of most Title IV-B sub-projects. Also they do not provide information on what specific skills and knowledge students do or do not have, and so they have a limited usefulness for planning instruction and remediation. For those Title IV-B sub-projects which have limited and well-defined objectives, then it will be most appropriate to have a criterion-referenced test with cut-off scores or standards representing acceptable levels of performance in the content area of interest. Commercially produced learning programs and textbooks often include such tests. When they are not provided, it will usually be necessary to rely on locally designed measures. In any case, it is not desirable to rely entirely on tests which accompany commercially produced materials. Instead one should develop additional more germane measures of local objectives and assess the extent to which the package contributes to the achievement of these same objectives. For those who need to develop their own measures, the bibliography at the end of this section suggests some useful sources on the construction of achievement tests. There are also numerous sources which coordinate and disseminate collections of objectives and test items appropriate for K-12 educational programs.

Pools of objectives and test items for grades K-12 in subject areas of language arts, math, science and social studies are available from:

Institute for Educational Research
100 West Maple Avenue
Downers Grove, Illinois 60515

The Co-Op
413 Hills House North
University of Massachusetts
Amherst, Massachusetts 01002

Instructional Objectives Exchange
(IOX Test Development Center)
Box 24093
Los Angeles, California 90024

Another source of valuable information on assessing achievement of knowledge and skills is the National Assessment of Educational Progress (NAEP). NAEP is a nonprofit project of the Education Commission of the States funded by the National Center for Education Statistics. NAEP conducts studies of the educational attainmens for 9 year-olds, 13 year-olds, and 17 year-olds across the United States in ten learning areas: art, career and occupational development, citizenship, literature, mathematics, music, reading, science, social studies, and writing.

Summaries of these studies as well as technical information on how local agencies may design and conduct local assessments as well as a free monthly newsletter on NAEP activities are available. Some of the NAEP test items are also made public and the results derived from tests based on national and regional samples can serve as a norm for interpreting local results. In the 1977-78 school year, an assessment was conducted in Connecticut using the goals, design, methodology, and some test items from NAEP, as well as test items that reflected state and local concerns. Information on the National and Connecticut assessments can be obtained from the following sources:

National Assessment of Educational Progress
Suite 700
1860 Lincoln Street
Denver, Colorado 80295

Connecticut Assessment of Educational Progress
State Department of Education
P.O. Box 2219
Hartford, Connecticut 06115

The following sources will provide introductions to various aspects of design and construction for achievement tests:

Bloom, B.S., Hastings, T.T., and Madaus, G.F. Handbook on Formative and Summative Evalution of Student Learning. New York: McGraw Hill, 1971.

Coffman, William E. Essay Examinations. In R.L. Thorndike (Ed.). Educational Measurements. Washington: American Council On Education, 1971.

Gronlund, N.E. Constructing Achievement Tests. Englewood Cliffs, New Jersey: Prentice Hall, 1977.

Gronlund, N.E. Preparing Criterion-Referenced Tests for Classroom Instruction. New York: The MacMillan Company, 1973.

Improving the Classroom Test: A Manual of Test Construction Procedures for the Classroom Teacher. Albany: State Education Department Bureau of Elementary and Secondary Educational Testing, 1973.

Mehrens, W.A. and Lehman, I.J. Measurement and Evaluation in Education and Psychology. New York: Holt, Rinehart and Winston, Inc., 1973.

Morris, L.L. and Fitz-Gibbon, C.T. How to Measure Achievment. Beverly Hills, Sage Publications, 1978.

Payne, D.A. Assessment of Learning: Cognitive and Affective. Lexington, Mass.: D.C. Heath and Co., 1974.

Wesman, A.G. Writing the Test Item. In R.L. Thorndike (ed.), Educational Measurement. Washington: American Council on Education, 1971.

Using Observation Procedures to Evaluate Changes in Student Behavior

Observation procedures of the type disscussed on pages 55-63 can also be used to evaluate changes in student behavior.

A set of performance objectives might be adopted related to the following desirable and undesirable social behaviors:
(a.) participating constructively in classroom presentations
(b.) Engaging in extended periods of work on curricular activities
(c.) Interferring with the work of fellow students
(d.) Ignoring instructions given by teachers.

Observations could be conducted to collect information on increases in the frequency of occurrence of the first two behaviors and decreases in frequency of occurrence of the two latter behaviors for sub-project participants.

Figure 20 shows baseline and criterion expectation levels for each of these behaviors. Figure 21 shows an instrument that might be used to tally occurences of these behaviors during regular classroom observations. Figure 22 shows a sheet that might be used to tabulate these results for groups of students and/or across the year for individual students. Figure 23 shows how a hypothetical set of results might appear for behavior (b) engaging in extended periods of work on curricular activities.

When it is expected that changes for individual or groups will not be dramatic enough to make clear judgements, then techniques for statistical analysis will usually be needed to support inferences concerning the stability and direction of changes in performance levels.

A more detailed discussion on observational procedures and a set of readings on this topic appears on pages 55-63.

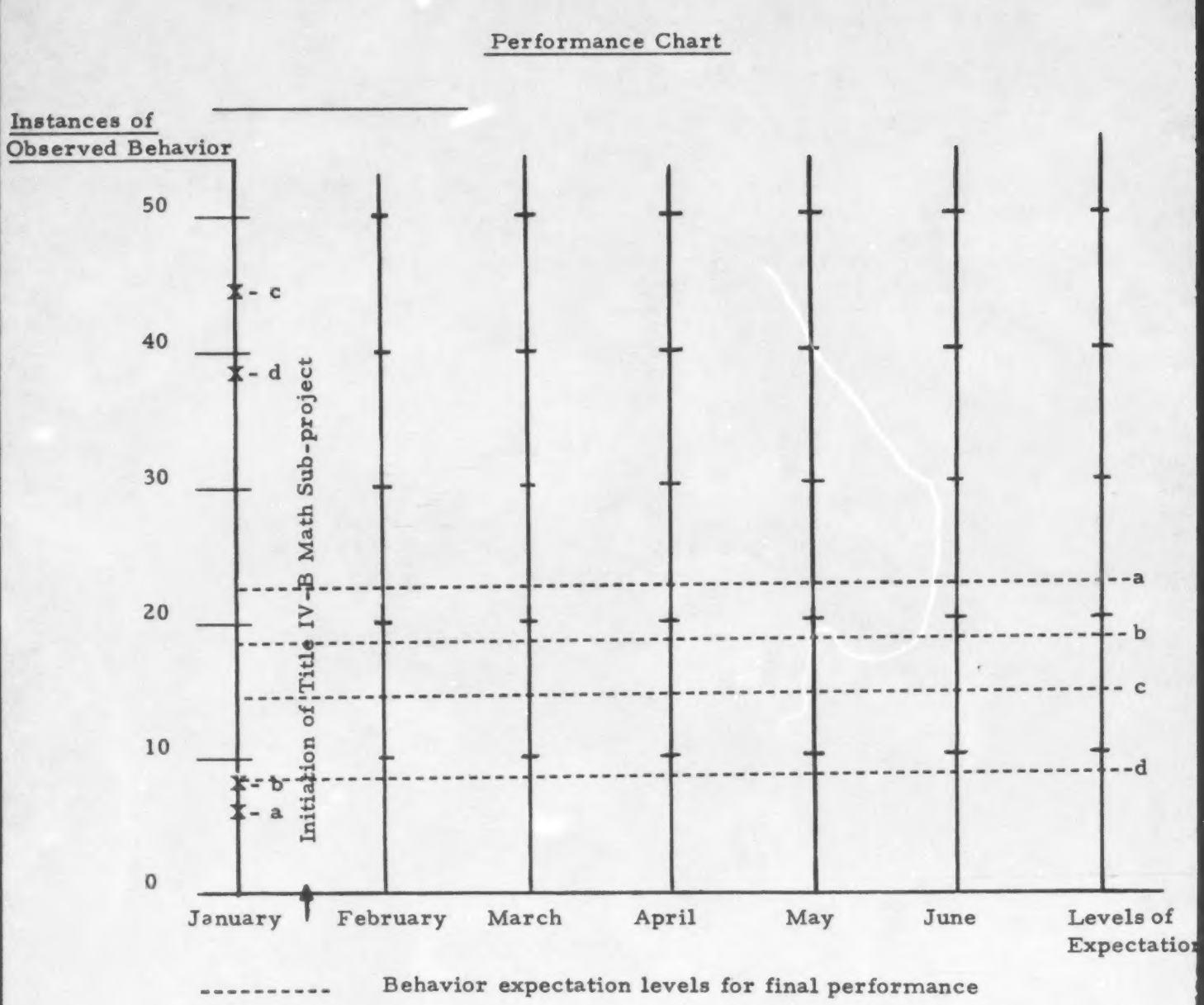


Figure 20. Base Line and Expectation Levels for Criterion Behaviors "a," "b," "c," and "d"

Behavior Observation Sheet

Name_____

Month_____

Number of Instances of Each Criterion Behavior

Observation Dates	a.	b.	c.	d.	Comments
1.					
2.					
3.					
4.					
5.					
6.					
7.					
8.					
Monthly Total					

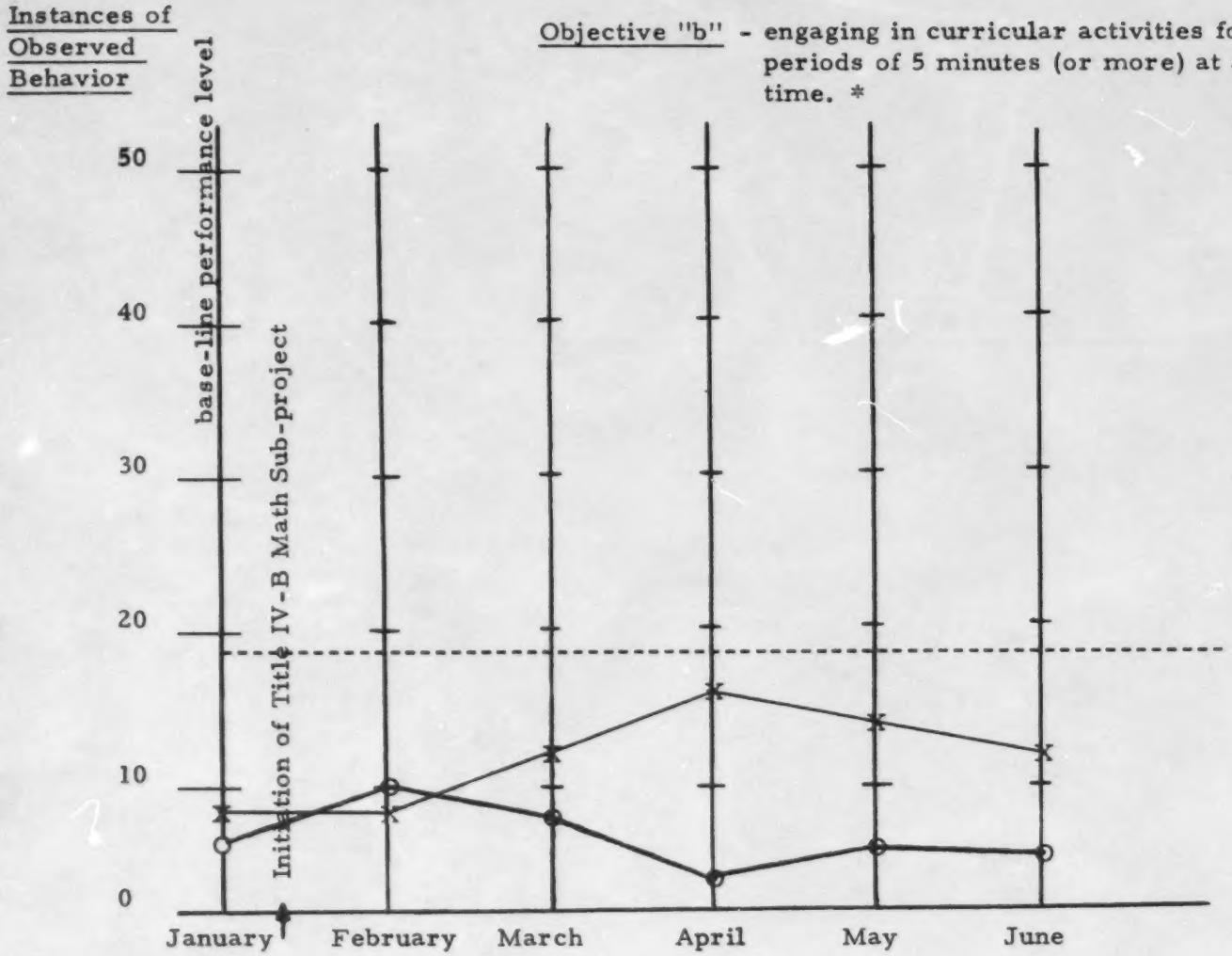
Figure 21. Classroom Observation Instrument for Tallying Instances of Criterion Behavior

Summary Sheet - Behavior "b"

Students	January	February	March	April	May	June	Total	Average
Total								
Average								

Figure 22. Tabulation Chart of a Given Criterion Behavior

Performance Chart



X - average performance levels for the group
O - monthly performance levels for ___(student's name)___
-------- expectation for final performance level - Behavior "b"

* Each period of 5 minutes counts as 1 "instance" (i.e., a 15-minute period of activity = 3 "instances")

Figure 23. Final Performance Levels on Criterion Behavior "b"

APPENDIX:
SUGGESTIONS FOR SELECTING FROM COMMERCIALLY PRODUCED INSTRUMENTS

Buros' Tests in Print describes, by subject area, all tests available in English-speaking countries. It is available in college libraries and large public libraries. In addition, test publishers will welcome inquires related to local testing and evaluation needs.

Best known publishers of educational tests:

Bobbs-Merril Company, Inc.
4300 West 62nd Street
Indianapolis, Indiana 46268

Bureau of Educational Research and Service
University of Iowa
Iowa City, Iowa 52240

Bureau of Publications
Teachers College Press
Teachers College
Columbia University
New York, New York 10027

California Test Bureau/McGraw Hill
Del Monte Research Park
Monterey, California 93920

Cooperative Test Division
Educational Testing Service
Princeton, New Jersey 93940

Educational and Industrial Testing Service
P.O. Box 7234
San Diego, California 92107

Educational Testing Bureau
Division of American Guidance Services, Inc.
720 Washington Avenue, S.E.
Minneapolis, Minnesota 55414

Harcourt Brave Jovanovich, Inc.
757 Third Avenue
New York, New York 10017

Houghton Mifflin Company
110 Tremont Street
Boston, Massachusetts 02107

Institute for Personality and Ability Testing
1602 Coronado Drive
Campaign, Illinois 61822

The Psychological Corporation
304 East 45th Street
New York, New York 10017

Psychometric Affiliates
Box 31167
Muster, Indiana 46321

Scholastic Testing Service,Inc.
430 Meyer Road
Bensenville, Illinois 60106

Science Research Associates Inc.
259 East Erie Street
Chicago, Illinois 60611

Stanford University Press
Stanford, California 94305

Organizations which will custom design tests for local use:

CTB/McGraw Hill (ORBIT)
Del Monte Reaearch Park
Monterey, California 93940

Houghton-Mifflin Company
(Customized Objective Monitoring Service)
777 California Avenue
Palo Alto, California 94304

Instructional Objectives Exchange
(IOX Test Development Service)
Box 24095
Los Angeles, California 90024

Multi-media Associates, Inc.
(Customized Criterion-Referenced Tests)
P.O. Box 13052
Tucson, Arizona 85732

National Evaluation Systems, Inc.
(CAM - Comprehensive Achievement Monitoring)
P.O. BOx 226
Amherst, Massachusetts 01002

Science Research Associates
(Mastery Custom Tests - Reading and Math)
155 North Wacker Drive
Chicago, Illinois 60606

Publishers will send specimen copies of tests and the accompanying test manuals. The following criteria will be helpful in evaluating tests for local use:

a) Consider appropriateness to local objectives. Ask appropriate members of the teaching staff to examine the test and comment on it. A rating scale may be used to elicit responses on cultural or sexual bias, age and content, appropriateness, and overall quality.

b) Examine the publisher's description of the test.

How was the test validated? Does this validation permit the kinds of inferences that need to be made. Is the sample on which the instrument was normed and validated sufficiently similar to the local district population?

What is the reliability of the test? Publishers of standardized tests should supply reliability coefficients or ranges of such coefficients for their tests. These coefficients indicate the extent to which the instrument performs dependably. A perfectly reliable test would receive a coefficient of 1.00 but this is rare. A reliability of .90 or more is considered quite good and a rating of .70 or less might suggest that the instrument is insufficiently reliable.

c) Consider the technical aspects of the test. How must it be administered or scored? Is this reasonable within temporal and financial constraints? In what form are the test results presented? Is this useful for local purposes? Two valuable aids in this process are Buros' Mental Measurements Yearbooks and the CSE Test Evaluation Scores edited by Hoepfner, Stein and Nummedal. The Yearbooks are a compendium of descriptions and critical reviews of published tests written by authorities in the field. The CSE series takes commercially available tests of student achievement and attitudes, organizes them by content areas (e.g., Foreign Language, mathematics), and rates them along dimensions of goal appropriateness, examinee appropriateness, technical quality and administrative usability.

d) The interpretation of test scores is a task which should be approached with great care. Reading some of the reviews in the Mental Measurements Yearbook will give an idea of the controversies that exist concerning what exactly is being measured by particular tests, even when these tests have gone through

elaborate phases of validation and standardization over many years. Caution then, should be exercised when interpreting an unreviewed or locally designed test. The modification of a published test for local purposes can be a useful and even a necessary procedure but again caution should be exercised because the interpretations and inferences permitted by the original instrument may no longer be warranted for the modified form.

The following sources can serve as an aid in finding and evaluting commercially produced tests:

Buros, O.K. Mental Measurements Yearbooks. Highland Park, New Jersey: Gryphon Press, 1972.

Buros, O.K. (Ed.) Tests in Print II. Highland Park, New Jersey: Gryphon Press, 1974.

Hoepfner, R. et. al. The CSE Test Evaluation Series. Los Angeles: Center for the Study of Evaluation, UCLA Graduate School of Education, 1971, 1976.

SOME FINAL SUGGESTIONS

An evaluation instrument should be selected as early as possible in the planning process. Ideally, in keeping with our evaluation principle of "last things first," concern for the evaluation method and instrument to be used should begin simultaneously with the setting of objectives in order to assure a maximum fit. If the evaluation is planned early enough in the planning process, expectations can be stated in a way that will facilitate evaluation. Likewise, laying out the analysis scheme ahead of time will help construct an instrument appropriate to the kind of information needed.

Evaluation is a cyclical process. This year's evaluation results can serve as the needs assessment for next year's sub-project and the baseline data for next year's evaluation effort. The planning phase of an evaluation is the point at which to assure that information collected will be useful for these purposes.

Just as it is important to avoid setting seemingly desirable but unmeasurable objectives, it is equally important to avoid the opposite extreme of limiting one's educational objectives just to what can be measured by an available test.

GENERAL BIBLIOGRAPHY ON EDUCATIONAL EVALUATION

Mehrens, W.A. and Lehmann, I.J. Measurement and Evaluation in Education and Psychology. New York: Holt, Rinehart and Winston, 1973.

Morris, L.L. (ed.). Program Evaluation Kit. Beverly Hills: Sage Publications, 1978. (This kit is composed of eight paperback volumes each covering a topic such as: goals and objectives, attitude measures, achievement measures, evaluation design, measuring program implementation and calculating statistics.)

Popham, W.J. Evaluation in Education: Currrent Applications. Berkeley: McCutchan, 1974.

Program Evaluator's Guide. (2nd ed.). Princeton, New Jersey: Education Testing Service, 1978.

Provus, M. Discrepancy Evaluation for Educational Program Improvement and Assessment. Berkeley, California: McCutchon, 1971.

Simplified Educational Assessment: A Manual of Non-Technical School Evaluation Techniques. The State Education Department. Bureau of School Programs Evaluation. Albany, New York, 1974.

Stake, Robert E. "The Countenance of Educational Evaluation." Teachers College Record. 68. 523-540, 1970.

Stufflebeam, D.L. et. al. Educational Evaluation and Decision-Making. Itasca, Ill.: F.E. Peacock, 1971.

Workbook on Program Evaluation. (2nd ed.) Princeton, New Jersey: Educational Testing Service, 1978.

07927

86